Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY · ORGANIZED IN 1935

VOLUME 16
NUMBER 3
SEPTEMBER
1951

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

SEPTEMBER 1951 VOLUME 16, NO. 3

Printed for the Psychometric Society at 23 West Colorado Avenue, Colorado Springs, Colorado. Entered as second class matter, September 17, 1940, at the Post Office of Colorado Springs, Colorado, under the act of March 3, 1879. Editorial Office, Department of Psychology, The University of North Carolina, Chapel Hill, North Carolina.

Subscription Price: The regular subscription rate is $10.00 per volume. The subscriber receives each issue as it comes out, and a second complete set for binding at the end of the year. All annual subscriptions start with the March issue and cover the calendar year. All back issues are available. The price is $1.25 per issue or $5.00 per volume (one set only). Members of the Psychometric Society pay annual dues of $5.00, of which $4.50 is in payment of a subscription to *Psychometrika*. Student members of the Psychometric Society pay annual dues of $3.00, of which $2.70 is in payment for the journal.

Application for membership and student membership in the Psychometric Society, together with a check for dues for the calendar year in which application is made, should be sent to

RAYMOND A. KATZELL, Chairman of the Membership Committee
Psychological Services Center, Syracuse University, Syracuse 10, New York

Payments: All bills and orders are payable in advance. Checks covering membership dues should be made payable to the *Psychometric Society*. Checks covering regular subscription to *Psychometrika* and back issue orders should be made payable to the *Psychometric Corporation*. All checks, notices of change of address, and business communications should be addressed to

ROBERT L. THORNDIKE, Treasurer, Psychometric Society and Psychometric Corporation
Teachers College, Columbia University
New York 27, New York

Articles on the following subjects are published in *Psychometrika*:

(1) the development of quantitative rationale for the solution of psychological problems;
(2) general theoretical articles on quantitative methodology in the social and biological sciences;
(3) new mathematical and statistical techniques for the evaluation of psychological data;
(4) aids in the application of statistical techniques, such as nomographs, tables, work-sheet layouts, forms, and apparatus;
(5) critiques or reviews of significant studies involving the use of quantitative techniques.

The emphasis is to be placed on articles of type (1), in so far as articles of this type are available.

In the selection of the articles to be printed in *Psychometrika*, an effort is made to obtain objectivity of choice. All manuscripts are received by one person, who

(*Continued on the back inside cover page*)

Psychometrika

CONTENTS

VOLUME SIXTEEN SEPTEMBER 1951 NUMBER THREE

NOTICE

The Educational Testing Service is offering for 1952-53 its fifth series of research fellowships in psychometrics leading to the Ph.D. degree at Princeton University. Open to men who are acceptable to the Graduate School of the University, the two fellowships each carry a stipend of $2,375 a year and are normally renewable.

Fellows will be engaged in part-time research in the general area of psychological measurement at the offices of the Educational Testing Service and will, in addition, carry a normal program of studies in the Graduate School. Competence in mathematics and psychology is a prerequisite for obtaining these fellowships. The closing date for completing applications is January 18, 1952. Information and application blanks will be available about November 1 and may be obtained from: Director of Psychometric Fellowship Program, Educational Testing Service, 20 Nassau Street, Princeton, New Jersey.

PSYCHOMETRIKA—VOL. 16, NO. 3
SEPTEMBER, 1951

A SPECIAL REVIEW OF *The American Soldier, Vol. IV**

PHILIP J. MCCARTHY
CORNELL UNIVERSITY

Volume IV of *The American Soldier* presents the first complete account of the scale analysis approach of Guttman and the latent structure approach of Lazarsfeld to the problem of attitude measurement. This review has been prepared for the purpose of providing an expository account of the models proposed by Guttman and Lazarsfeld, together with an indication of the places which call for additional clarification and research.

1. *Introduction*

This paper will present an expository review of Volume IV of *The American Soldier*. The materials in this volume are divided into two more or less distinct parts. The first eleven chapters deal with theoretical and empirical analyses of problems of measurement—in particular, the scale analysis approach of Louis Guttman and the latent structure approach of Paul Lazarsfeld. The last five chapters are devoted to an account of two specific studies in prediction, namely, the screening of psychoneurotics in the army and the postwar plans of soldiers. This review will be focussed upon the contents of the first eleven chapters.

Scale analysis and latent structure analysis are of fundamental importance because they provide a conceptual framework (or, a model) with which to attack the problem of attitude measurement. In particular, they attempt to test the hypothesis that a delimited area of human behavior (which may refer to observed reactions to specific situations, or to verbal expressions of feeling toward specific situations, and so on) contains only a single dimension (the hypothesis of unidimensionality). If the hypothesis of unidimensionality is accepted, then one can consider the task of arranging people in rank order with respect to this single dimension. This does not necessarily mean that ordering cannot be carried out if the hypotheses of unidimensionality is rejected—only that the ordering is more meaning-

*This review was originally prepared at the request of the Sociological Research Association, and was presented at the Association's Annual Meeting in New York City, December, 1949.

ful if a single dimension is present. Volume IV of *The American Soldier* gives the first complete account of the work of Guttman and Lazarsfeld. This work will not solve all problems of attitude measurement. However, it will permit other research workers to examine the models which have been set up, to test them out empirically in a wide range of situations, and to change or discard them as empirical evidence grows.

Whenever one sets up a model or theoretical framework for describing a particular phenomenon, the following questions arise more or less naturally:

(1) Does the model at least provide a logical description of the phenomenon under study?
(2) Can the model be subjected to objective verification?
(3) When the model holds, does it permit one to make important and useful deductions about the phenomenon?
(4) Does the model fit a wide enough range of cases to make it practically useful?

It is with respect to these four questions that scale and latent structure analysis will now be examined.

2. *The Model for Scale Analysis*

(a) *The approach.* Guttman advocates considering an attitude as a *delimited totality of behavior with respect to something.* He does not take this as a complete definition, but only as a necessary component in a definition. Moreover, it is a component which can, hopefully, be given an operational meaning. Once this operational meaning has been discovered and observed, then one can attempt to weave the results back into a social-psychological definition of an attitude.

The notion of an attitude as a totality of behavior toward something requires that the population of individuals whose behavior is to be observed is specified. The moment in time at which the observed behavior occurs is a necessary part of the population definition. Another moment in time may show a different pattern of behavior. In this section it will be assumed that the behavior of the entire population is observed. In practice, it will usually be necessary to study only a sample from the population and this aspect of the problem will be touched upon in the next section.

Having defined the population of individuals with respect to which the attitude is to be defined, the next implication of the above definition is that the behavior of each population element toward "something" must be observed. What is this "something" to be? Actually, this can be interpreted very broadly. Each individual may be

placed in a series of concrete situations, and his reactions to each situation may be observed; or each individual may be asked a series of questions, and his answer to each question may be recorded. This latter case is the one which ordinarily occurs in attitude research, and it is the only one in which scale analysis has thus far been applied to any appreciable extent. Narrowing the range of "something" to the asking of questions then leaves the problem—what questions?

Although it has not been explicitly brought out in the preceding paragraphs, one is usually in the position of being able to state, at least in broad terms, the general nature of the attitude under investigation. It may be attitude toward Russia, attitude toward the United Nations, attitude toward socialized medicine, or attitude toward low-cost housing. However, whatever the specific topic may be, there will always exist (at least conceptually) an indefinitely large number of questions which might be asked. The aggregate of all these questions is termed by Guttman the *universe of content,* or the *universe of attributes.* Here, just as for individuals, the universe must be sampled in order to carry out a practical study; and again the sampling aspects will be relegated to the later portions of the review. For the moment, the existence of this universe of content will be taken for granted.

Ideally, the answer of each person in the population to each question in the universe of content is obtained and recorded. These responses can then be examined as they relate to the following two fundamental problems of attitude measurement:

(1) Does the pattern give evidence concerning the hypothesis of unidimensionality?

(2) If only one dimension is present, can people be given a unique rank order with respect to favorableness or unfavorableness on the attitude?

Guttman proposes one particular pattern of responses—namely, the scale pattern—for which it seems to be possible to give an unequivocal answer to these questions.

(b) *The model or scale pattern.* Assume that the universe of content is made up of questions, each of which consists of a series of categories graded from favorable to unfavorable (or vice versa). It is not necessary that the number of categories be limited to two (i.e., a dichotomy). A person responds to a question by placing himself in the category which most nearly mirrors his position. Thus in studying the attitude of enlisted men toward their officers, such questions as the following were asked:

"How do you feel about the privileges that officers get compared with those which enlisted men get?

1. ————Officers have *far too many privileges*
2. ————Officers have a *few too many privileges*
3. ————Officers have *about the right number of privileges*
4. ————Officers have *too few privileges*"

Under these circumstances, it seems logical to state that one individual from the population will have a higher ranking in respect to the attitude than another individual if the first is just as *high* or *higher* on every question or item in the universe of content than is the second. As a matter of fact, this is the definition of a scale. To quote: "We shall call a set of items of common content a scale if a person with a higher rank than another person is just as high or higher on every item than the other person."

The above definition of a scale immediately leads to the parallelogram response pattern. Consider a universe of three questions A, B, and C, each of which has three categories graded from favorable to unfavorable (A_1, A_2, A_3; B_1, B_2, B_3; C_1, C_2, C_3). It is assumed that each person in the population places himself in one category for each of the three questions. If a scale exists in the sense of the preceding definition, there are only seven possible response patterns. These are (assuming the smallest proportion endorse A_1, the next smallest B_1, and the next smallest C_1):

Rank of Individual	Categories								
	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
6	x	x	x						
5		x	x	x					
4			x	x	x				
3				x	x	x			
2					x	x	x		
1						x	x	x	
0							x	x	x

It might be noted that if the questions have more than two categories, one can obtain slight deviations from the perfect parallelogram pattern and still have the scale criterion satisfied. That is, the parallelogram pattern is nothing more than a visual aid in explaining scale theory. In the parallelogram pictured above, the reaction of the entire population to the entire universe of content can be completely specified by giving the proportion of people having rank 6, the proportion having rank 5, and so on. Notice particularly that knowledge

of a person's rank immediately tells his pattern of responses. It is extremely easy to write down information and knowledge questions which one would expect to scale, and many simple illustrations of this type are given throughout Volume IV.

The high degree of consistency required of the response pattern demands that the internal ordering of the categories in each question be with respect to some common element. Guttman refers to this by saying that all the items have a single content meaning. Moreover, the rank order of an individual contains all of the information which would be available if the responses to each of the questions were kept distinct. To put this in statistical terms, rank order is a *sufficient statistic*. These observations arise through a purely formal analysis and it would be a mistake at this stage to read more meaning into them than actually exists. The scale itself does not define or name content any more than a correlation coefficient implies cause and effect. It is relative to the population (composition and time of questioning) and to the defined universe of content. A broader universe of content might or might not give a scale (any subset of the original universe will always give a scale). All of these points must eventually be given a psychological meaning, perhaps by tying this formal analysis back into a social-psychological definition of an attitude.

One final comment on the scale or parallelogram pattern concerns the nature of items in the universe of content. It has already been noted that each item must contain two or more categories which can be naturally ranked from favorable to unfavorable. If a scale is to exist, these individual items must be essentially cumulative in nature. The prototype of items having an intrinsic cumulative character is the social distance scale, with such items as the following (example from Chapter 1):

1. Would you want a relative of yours to marry a Negro? ("Yes" or "No").
2. Would you invite a Negro to dinner at your home? ("Yes or "No").
3. Would you allow a Negro to vote? ("Yes" or "No").

It is possible to ask questions which do not have this cumulative character, even though a single content meaning is still present. Under appropriate circumstances, this can lead to a pattern of responses different from the parallelogram pattern, where there still exists a one-to-one relationship between rank order and response pattern. This problem is discussed in Chapter 1 and undoubtedly deserves further investigation.

(c) *Inferences drawn from the scale model.* If a population of individuals and a universe of attributes produce a perfect scale, there are certain useful conclusions which can be drawn, either by a logical examination of the scale pattern or by the use of mathematical analysis. A brief account will now be given of the more important of these.

Perhaps the most important aspect of the scale pattern is the simple fact that a person's responses to every item in the universe can be reproduced from a knowledge of his rank alone. Not only does this make the description of the "totality of behavior" a simple one, but it also means that a person's rank order exhausts or summarizes the information which all items can give about him. Consequently, the conclusion which is stated so frequently in Vol. IV follows immediately: "The zero order correlation with the scale score is equivalent to the multiple correlation with the universe." Guttman refers to this as the problem of *external prediction.* The above quotation implies that one is interested in a linear combination of the questions as a predictor. Rank order might not be the "best" predictor if curvilinear regression were more appropriate for predicting a particular external variable.

The second generally useful result flowing out of the existence of a scale is the manner in which it allows one to attack the problem of drawing a sample of questions from the universe of content. If a perfect scale exists, then one can be sure that any sample of questions selected from the universe will scale (i.e., form a cumulative or parallelogram pattern) and that people will be placed in their proper rank order. Individuals having the same rank in the sample of items might have different ranks if all items were used—and probably would. However, a person having a higher sample rank than a second person would, of necessity, have a higher rank on the entire universe. In theory at least (and practice will be discussed in the next section), the existence of a scale solves the question-sampling problem. One does not need to use the entire universe, but only enough questions to provide the desired number of scale ranks.

As noted above, an individual's rank order may be all that is needed for external prediction. There is no need to worry about scores for people or weights for the categories of the various items. However, it is sometimes necessary to inquire a bit more deeply into the internal relationships (i.e., within the scale pattern) between persons having different ranks, between different categories of the same question, and between different questions. Considerations of this kind introduce the problem of a metric. In other words, can individuals be

given a more meaningful quantitative score (on the scale from favorable to unfavorable) than their rank order?

Guttman has approached this question through the well-known principle of least squares. Quoting from Chapter 9: "The most internally consistent scores to assign the people on the basis of their responses to the items are those that satisfy the following condition. All people who fall in one category of an item should have scores as similar as possible among themselves, and as different as possible from the scores of the people in the other categories of the item; this should be true to the best extent possible for all items simultaneously." These scores can be obtained without introducing the notion of category weights, and it is then possible to determine independently the most consistent category weights through the use of a criterion similar to that set forth above. However, it can be proved that there is a relationship between the "best"scores and the "best" weights. Briefly, the score of a person is proportional to the arithmetic mean of the weights of the categories by which he is characterized, and the weight of a category is proportional to the arithmetic mean of the scores of the people who are in it. As far as present applications of scale analysis are concerned, this fact seems to be used only as a justification for the simple scoring system which is ordinarily applied. However, it is of some theoretical interest to have this "duality" between persons and questions.

In addition to justifying the use of a simple scoring system in the applications of scale analysis, the search for a metric also introduced the concept of principal components. It was found that there were more than one set of internally consistent scores, that is, scores from which weights could be derived which would give back the original scores (at least within a factor of proportionality). As a matter of fact, if the universe of attributes contains m different types of dichotomous items (two items being of the same type if they are perfectly correlated), then there will be m different sets of scores. However, these various sets of scores will differ in respect to their degree of internal consistency (see criterion above). This means that one now has three different ways of characterizing the way in which a person reacts to the universe of attributes, namely, (1) by his response pattern, (2) by his rank, or (3) by his score on *each* of the principal components. The third characterization will perhaps prove to be the most helpful in attempting to tie scale analysis back into a social-psychological definition of attitude. Although a few advances in this direction have been made, most of this must wait on future research.

The most internally consistent set of scores has been called the "content" scores. It can be shown that an individual having a higher rank order than a second will also have a higher content score, and vice versa. There seems to be little doubt that this stands up under logical examination. The second most internally consistent set of scores (the second component) has been shown on a purely formal basis to be a U-shaped function when plotted against rank order. Experimental work indicates that "intensity of feeling" for attitude questions has a U-shaped relationship with rank order. In other words, the less favorable a person is on a certain attitude the more intensely he holds this position, and similarly for persons on the favorable end of the scale. The third component, considered as a function of rank order, will have two bends, and so on through all the components which exist in a particular universe of content. Each of these components is a perfect (but curvilinear) function of rank order. However, they all have zero linear correlations with one another.

All of the above outlined characteristics of the principal components are illustrated from a formal point of view in Chapters 1 and 9 of Volume IV. In addition, the application of the second or intensity component to the work of the Research Branch is described. It would appear at the present time that there may be some hope in the future of isolating, defining, and measuring independently components above the second. As a matter of fact, Guttman, in a personal communication, has indicated that he has succeeded in empirical isolation of the third component. This has been called "closure." However, there is no guarantee that this process can be carried on indefinitely, or even that there is any necessary correspondence between the mathematical components and the empirically determined psychological components. There need be no psychological equivalent of the quantities derived solely from mathematical analysis. This mathematical analysis can be taken as a guide for future research, but it cannot replace it. Furthermore, there are many fundamental aspects of the model itself which need to be investigated ("how to define universe of content?" and "how frequently does the model apply?") before attention is devoted solely to the "overtones" of the principal components.

3. *The Scale Model in Practice*

(a) *Deviations from a perfect scale.* Thus far no populations and universes of content, at least insofar as the measurement of attitudes is concerned, have been found which give rise to a *perfect* scale.

This is not surprising if one only remembers the highly restrictive nature of the scale pattern. Notice that this observation does not depend upon studying the entire population and the entire universe of content. If a perfect scale exists, then *any* sample of individuals and *any* sample of items must also show the scale pattern.

The first question that one may ask concerns the reasons for obtaining deviations from a perfect scale. This can only occur because there is more than a single dimension operative in the reaction of individuals to questions. This may imply that there are several major "content meanings" inherent in the universe of content, that individuals are not consistent in their reaction to the questions (the problem of test-retest reliability), or that there are many minor variables causing disturbances. The contents of Volume IV admittedly treat these problems from a more or less intuitive point of view. Until a body of experience has grown up, this may be all that one can do. However, ultimately a more precise explanation must be given.

At this stage in the development of scale theory, the classification of a deviant scale pattern into one of the above categories is based mainly on visual inspection. Thus, if the proportion of errors is small (and a discussion of "small" will be given shortly), and if they are scattered randomly through the various rank orders, an approximate scale is said to exist. In other words, it is assumed that many small variables—unimportant for descriptive or predictive purposes—are causing the disturbances. On the other hand, if the errors are relatively large and tend to be concentrated in a particular rank group, it is inferred that a major variable is distorting the pattern. The final type differentiated is the case where the errors occur most frequently in the midrank range, less frequently in the extreme rank individuals, and are distributed randomly among individuals and categories. This is referred to as a *quasi-scale.* Illustrations of these three types are given in Chapter 5. Here, as always, a more precise definition of randomness is required. However, even with such a definition, these three categories will tend to blend into one another and a sharp distinction will probably never be possible.

Concurrently with the types of deviations which may occur, one must also be interested in the amount of error which occurs. There are several different ways in which this error can be measured, two of which can be easily illustrated with the following pattern.

RESPONSE PATTERNS OF INDIVIDUALS

Individuals	Categories								
	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
10	x	x	x						
9	x	x	x						
8		x	x				(x)		
7		x	x				(x)		
6			x	x	x				
5			x	x	x				
4	(x)				x	x			
3				x	x	x			
2					x		x		(x)
1							x	x	x

The first and most obvious way to measure error is to compute the proportion of non-scale individuals. For the above pattern, this is clearly 4/10 or 40 per cent. The difficulty with this measure is that it does not take into account the number of items in the universe of content. Thus an individual is considered as one deviant whether his responses to all questions deviate from the scale pattern or whether only one response out of a large number deviates from the scale pattern. In this instance there are four deviant responses (enclosed in parentheses in the above diagram) out of a total of 30 responses—13 per cent error. If we take one minus the proportion of error measured in this way, we obtain Guttman's *coefficient of reproducibility*. It would be possible to make still further refinements in the measurement of error if one considered not only the deviant responses but also the number of categories by which they deviate from the position they would occupy in a perfect scale. However, this would involve some idea of "distance" between categories. Also, it is not always possible to tell exactly which of the responses are the deviant ones. All of this discussion is still in terms of the entire population and the entire universe of content. The sampling problem will come in the next subsection.

Since perfect scales are not likely to exist, much less arise in practice, how much error can be tolerated before the theoretical advantages of scale analysis begin to break down? It is at this point that Volume IV leaves one more or less with the impression of being suspended in mid-air. The following two guides are given:

(1) If the *population* coefficient of reproducibility is in the neighborhood of .90 (and the errors are random), then it is implied that one will gain all the advantages of having a perfect scale (ranks will adequately predict responses, ranks will serve for external pre-

diction, and internal analyses for a metric, intensity and other components will be valid).

(2) No matter how low the *population* coefficient of reproducibility may be, if the errors are distributed in the random gradient pattern of a quasi-scale then the ranks (which will no longer adequately reproduce responses) will still provide the best means of external linear prediction. This implies that there is a major single content variable, but that it is disturbed by many small random effects.

It is recognized that the critical coefficient of reproducibility must be set somewhat arbitrarily at the present time, but there must have been some fortunate or unfortunate experiences which led to the choice of .90 as the critical value. An account of this previous experience might serve as a useful guide in future research.

(b) *Testing the hypothesis of approximate scalability.* The fundamental problem in applying scale analysis to any particular situation lies in determining whether or not an approximate scale actually exists. Once it has been determined that an approximate scale exists, the advantages set forth in the preceding section follow. The testing procedure depends upon defining the population of individuals and drawing a sample from that population, upon delimiting the universe of content and selecting a sample of items, upon observing the reaction of the individuals to the sample of items, upon computing the coefficient of reproducibility from the samples, and finally, upon inferring from the sample results what would have arisen had the entire population reacted to the entire universe of content.

The definition and sampling of the population of individuals is not particularly different for scale analysis than for any other research problem since one desires to draw valid inferences from the sample concerning the entire population.

The definition and sampling of the universe of content is a much more formidable task than is the sampling of people. Scale analysis cannot define content. It can only tell what to do with content after a scale has been shown to exist. The materials of Volume IV are not particularly helpful in this respect. It is all very well to state that one should first give the universe of content a name, should then write down all questions which seem to relate to this universe, and should then select a sample with which to test scalability. Although this may sound like a simple procedure, it is, in reality, a very complex one. This is clearly stated in Volume IV, and although some general rules are given to ensure that a spuriously high coefficient of reproducibility will not be obtained in the testing procedure (see succeeding

paragraphs), the major tasks still remain. Some advances in this direction have already been made by combining the Thurstone, Likert, and Guttman approaches to scale construction by Edwards and Kilpatrick (1).

Once the sample of people has been selected and the sample of questions has been prepared, one enters the practical phase of asking the questions. There has been some doubt raised concerning the feasibility of maintaining rapport with respondents when many questions bearing on the same content are being used. This has been specifically raised by Festinger (2), but the real test of this point can come only through experience. This reviewer has had no experience on this score. Neither has he had an opportunity to review the literature and summarize others' experience. Many successful applications are reported in Volume IV, but it can be argued that the polling of soldiers in an army atmosphere is not indicative of what will happen with other populations.

After the answers to the questions have been obtained, the sample coefficient of reproducibility must be determined. This is not a trivial problem since it is necessary to determine, on a more or less "cut and try" basis, that ordering of individuals and that ordering of questions which will give the highest reproducibility. Two chapters in Volume IV, prepared by Suchman, describe in great detail the use of a "scalogram" board for carrying out this operation. This is a very neat mechanical device for doing a difficult job, but its use is handicapped by the fact that only about 100 individuals can be treated at one time. A paper and pencil technique, which suffers from this same restriction as to size of sample, has been described by Guttman (4). Recently, Ford (3) has suggested a procedure which makes use of IBM punched card equipment and which therefore imposes no restrictions on the number of individuals used in the testing. The only troublesome feature of these approaches lies in the combination of categories of various questions. This is justified on the grounds that one cannot tell in advance which categories will actually be distinct to the respondents. There does not seem to be any serious theoretical difficulty inherent in this combination of categories, provided some of the general rules now to be set forth are observed.

When the individuals are arranged in rank order, the preceding work gives rise to a pattern of responses and to a coefficient of reproducibility. From these two items of information, one must infer what would have happened had the entire population reacted to the entire universe of content. This inference, at the present time, rests mainly

on certain "rules of thumb" which appear adequate until more precise rules can be determined.

The primary concern is that the sample coefficient of reproducibility is not *spuriously* near .90—i.e., near .90 solely as a result of chance. In order to give an exact answer (in terms of probability) to this question, much more needs to be known about the entire process than is known at present, and in particular, sampling theory for the reproducibility coefficient is required. However, Vol. IV presents some rules to follow which will help out in this respect. These are:

(1). A sample of at least 100 individuals should be used in testing the hypothesis of approximate scalability.

(2) The more items included in the test, the greater is the assurance that the universe is scalable. A sample of at least ten items is recommended for this task.

(3) The more response categories for items included in a test, the greater is the assurance that the entire universe is scalable.

(4) If all sample items have marginal distributions in the neighborhood of 80-20 or 90-10 splits (i.e., for dichotomous items), then spuriously high coefficients are likely to be obtained. Therefore, it is recommended that the sample should contain items having as wide a range of marginal distributions as possible, and specifically, it should include items with marginals around 50-50. Moreover, items should have at least 90 per cent reproducibility by themselves, or should contain "more non-error than error."

(5) The pattern of errors should be examined to see that there are no groupings of non-scale types. This problem of pattern has already been treated eariler in this review.

These criteria are discussed and illustrated at some length in Vol. IV with both empirical data and hypothetical examples. Lacking a theory of sampling for the reproducibility coefficient, they are perhaps the best that one can do. Experience and research will tell more about this.

(c) *Measuring intensity and the zero point.* It is shown empirically in Chapter 7 that intensity of feeling bears the U-shaped relationship to rank order that theory predicts for the second component. Consequently, the second component has been given the name of "intensity." The low point of the "U" (or the zero point) is interpreted as a cutting point on the scale to separate the "favorables" from the "unfavorables." Unfortunately, measurement techniques are not available for making intensity the *perfect* function of rank order that theory predicts. Much error is inherent in its measurement.

One cannot doubt the usefulness of a "zero point." However, there are many questions concerning the interpretation of intensity as the second component. Is intensity really the second component or does it only appear so because of the large amount of error involved in its measurement? If intensity actually is the second component in some instances, is it always the second component? May not the identity of the second component vary from one type of situation to another? Does the cutting point provided by the second component actually divide the population of individuals into "favorable" and "unfavorable," or must some other interpretation be given to this cutting point? The answers to these questions would seem to be dependent at least partially on finding improved methods of measuring intensity and on applying these and the present methods to a wide range of practical situations.

(d) *The incidence of scales.* The amount of effort which should be expended on the unsolved problems of scale analysis must be determined somewhat from the extent to which they are applicable in practical problems. It is admitted in Volume IV that the occurrence of a scale is probably the exception rather than the rule. Many examples of scalable universes and populations are given, and many more examples are cited where scalability was not found. At this stage in development, it would seem very worthwhile to make a survey of attempted applications and to summarize this information. This reviewer would have liked very much to have included such a survey, but time was not available.

4. *The Model for Latent Structure Analysis and its Relationship to the Scale Model*

(a) *Preliminary remarks.* It would appear likely that any "uninitiated" person will be left with a feeling of frustration after a first reading of Chapters 10 and 11 on latent structure analysis. Such a situation will be unfortunate, but also perhaps unavoidable. The difficulties stem from three facts. These are:

1. Scale analysis attempts to pick out a particular pattern of responses (of people to questions) and to extract from this pattern of responses all of the pertinent information. The basic approach is simple, the ideas are simple, and much experience has been developed. On the other hand, latent structure analysis attempts to set up a model which will be applicable to a much wider range of response patterns and the underlying concepts are as a result more complex.

2. Scale analysis, in theory and in practice, can be presented

with a minimum amount of mathematical analysis (except where principal components are involved) while latent structure analysis, once the underlying model has been set up, depends very heavily on mathematical formalization and its application requires a large amount of numerical computation.

3. Scale analysis is spread over eight chapters, while latent structure theory is concentrated in two. This is a direct result of the fact that scale analysis developed and was applied over the entire period of World War II, whereas latent structure theory grew up in a short space of time and has not as yet had a great deal of practical application.

The above points have been set forth for two reasons. First, it is to be hoped that a clear understanding of their implications will better enable readers of Volume IV to pick out the salient features of latent structure theory without being repelled by the complexity of its mathematical and computational aspects. In this respect it might be noted that a very well written, general account of latent structure theory is presented in Chapter 1 (by Stouffer). It is strongly recommended that a general reader examine this material very carefully before proceeding to Chapters 10 and 11. The second reason for making these observations is so that one can better realize the limitations which are imposed on a review such as the present one. The wide ramifications of latent structure theory simply will not admit as clear an exposition as did the rather narrower concepts of scale analysis.

(b) *The model.* Although it is not explicitly stated in Volume IV, and much confusion would be avoided if it were, latent structure theory starts at the same point as does scale theory. A population of individuals and a universe of questions (of common content) is assumed. Thus far the main body of theory has been developed for dichotomous questions. The reaction of each individual to each question is observed and the results are called the "manifest data." One can now state (quoting from Chapter 1) that "The latent structure approach is a generalization of Spearman-Thurstone factor analysis. The basic postulate is that there exists a set of *latent classes, such that the manifest relationship between any two or more items on a questionnaire can be accounted for by the existence of these latent classes and by these alone.* This implies that any item has two components—one of which is associated with latent classes and one of which is specific to the item. The specific component of any item is assumed to be independent of the latent classes and *also independent*

of the specific component of any other item." Actually, at this stage it would be better to say not "items on a questionnaire," but "items selected from the universe of content."

Before proceeding further, one can see at a glance that this postulates a much more complex situation than did the Guttman criterion that responses should be reproducible from the ranks (i.e., of individuals). Moreover, the parallelogram pattern of responses of scale analysis satisfies the latent structure postulate, the ranks defining the classes of the structure.

Latent structure theory attempts to explain the observed response pattern in terms of an *underlying* attribute. Or, to put it in different terms, it assumes that there is a variable (according to which each individual can theoretically be classified into one of a set of mutually exclusive classes) which accounts for all the observed relationships in the manifest data. These latent classes can be conceptualized in many ways. For example, Chapter 10 starts with a continuous variable, x, and derives the latent classes by partitioning the range over which x varies. In this case, the latent classes have a natural ordering; but in other cases, such a natural ordering may not be apparent from the model or from the computations made with the manifest data.

In the discussion of scale theory it was observed that a scale could not define the content. It could only show that there was a single content meaning, or a single dimension, existent in the individual question reaction. Just so, the existence of a set of latent classes cannot, by itself, provide a name for the latent attribute or for its classes. This situation is strictly analogous to that existing in factor analysis.

The general model hypothesized by latent structure theory is as follows. Suppose that there is a latent attribute having λ classes and that there are m dichotomous questions in the universe of content. Let $n_I, n_{II}, \cdots, n_\lambda$ be the number of people falling in the respective classes of the attribute. Furthermore, let p_{ij} be the proportion of people in the ith class ($i = I, II, \cdots, \lambda$) who answer "Yes" to the jth question ($j = 1, 2, \cdots, m$). In tabular form, this is

Latent Class	Latent Class Frequency	Items			
		1	2	. . .	m
I	n_I	p_{II1}	p_{II2}		p_{IIm}
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
λ	n_λ	$p_{\lambda 1}$	$p_{\lambda 2}$		$p_{\lambda m}$

The total number of individuals represented by this table is n (i.e., $n_I + n_{II} + \cdots + n_\lambda = n$). The requirement that all relationships among the items be accounted for by the latent classes alone can be translated into the statement that items should be independent within classes. This leads to a set of equations of which the following two equations are examples:

$$n_{12} = n_I\, p_{I1}\, p_{12} + n_{II}\, p_{II1}\, p_{II2} + \cdots + n_\lambda\, p_{\lambda 1}\, p_{\lambda 2}\,,$$
$$n_{123} = n_I\, p_{I1}\, p_{12}\, p_{13} + n_{II}\, p_{II1}\, p_{II2}\, p_{II3} + \cdots + n_\lambda\, p_{\lambda 1}\, p_{\lambda 2}\, p_{\lambda 3}\,.$$

n_{12} is the number of persons giving "Yes" responses to both items 1 and 2, and n_{123} is the number of persons giving "Yes" responses to all three items 1, 2, and 3. These, and the other equations which must hold, state that the responses of people falling in the same class of the latent attribute to any set of questions selected from the universe of content must show independence between the questions. Lazarsfeld puts this intuitively by saying that the latent attribute is the only thing which holds the questions together. In practice, one will not know that a latent attribute exists; and if it does exist, one will not know the latent structure parameters $n_I\,, n_{II}\,, \cdots, n_\lambda\,, p_{I1}\,, p_{12}\,, \cdots, p_{II1}\,, p_{II2}\,, \cdots, p_{\lambda 1}\,, \cdots, p_{\lambda m}$. These parameters must be determined from the manifest responses, once it has been ascertained that the latent structure model holds. Further discussion of these points will be given under the section on application.

(c) *Deductions following from the latent structure model.* Assuming that the latent structure model is satisfied by a population of individuals and by a universe of content, one may well ask what useful deductions follow. Perhaps the foremost problem in this respect is that of sampling, for no practical applications can be made without induction from a sample. The general aspects of this situation are barely mentioned in Volume IV. However, the following comments can be made. If the entire population of individuals is used, then no matter how a sample of questions is chosen from the universe of content, the latent structure model must be satisfied. Furthermore, the same type of statement seems to hold true for the sampling of individuals. Note, however, that the structure parameters $n_I\,, n_{II}\,, \cdots, n_\lambda$ (or more appropriately, when a sample is being considered, $n_I/n\,, n_{II}/n\,, \cdots$) will depend very markedly on the way the population of individuals is sampled. In other words, a "good" sample of individuals is required. This is not essentially different from the determination of the proportion of people having various scale ranks as dis-

cussed in the portion of this review devoted to scale theory.

The structure parameter p_{ij} was defined as the proportion of people in the ith class who answer "Yes" to the jth question. An alternative definition might be: p_{ij} is the probability that a person of class i will answer "Yes" to the jth question. The use of this latter definition would essentially introduce the idea of test-retest reliability into the model. Still a third interpretation might be given to p_{ij} by assuming that it is the *average* probability that a person in class i will answer "Yes" to the jth question. Some thought should be given as to the relative merits of these three alternative definitions, and to the changes which they might introduce into the latent structure theory.

At the present time, the latent structure model would seem to be most useful where it provides a natural means of ranking latent classes and response patterns. The simplest case of this is that of a latent dichotomy (i.e., each individual either possesses the attribute or does not possess the attribute). In this instance, response categories can be ranked according to the proportion of individuals in that category who possess the attribute. For example, suppose that there are three items in the universe which give rise to a latent dichotomy. Then it is possible to compute for each of the eight response patterns (+++, ++−, −++, +−+, +−−, −+−, −−+, −−−), where a "+" signifies a "Yes" and a "−" signifies a "No," the proportion of individuals having that pattern who possess the attribute. This provides a means of ranking response patterns (except, of course, that one cannot tell within a specified response pattern which individuals possess the attribute and which do not possess it).

Certain other latent structures which have this special order feature are picked out for special attention in Chapters 10 and 11. For example, a Guttman scale has the following structure:

GUTTMAN SCALE

Latent Class Frequency	Items						
	1	2	3	4	5	...	m
n_I	1	1	1	1	1	...	1
n_{II}	0	1	1	1	1	...	1
n_{III}	0	0	1	1	1	...	1
.							.
.							.
.							.
n_λ	0	0	0	0	0	...	0

A quasi-scale would seem to have the following latent structure:

QUASI-SCALE

Latent Class Frequency	Items 1	2	3	...	m
n_I	$1-a_1$	$1-a_2$	$1-a_3$	...	$1-a_m$
n_{II}	a_1	$1-a_2$	$1-a_3$	...	$1-a_m$
n_{III}	a_1	a_2	$1-a_3$	...	$1-a_m$
.					.
.					.
.					.
n_λ	a_1	a_2	a_3	...	a_m

In this case it is assumed that $a_1, a_2, \cdots, a_m$ are "small" probabilities.

The concept of a natural ordering is generalized by Lazarsfeld to cases which he defines as possessing *latent linearity*. However, there is neither time nor space available for discussing this in detail. At the present time it would appear to be difficult to talk meaningfully about any kind of latent structure which does not possess some kind of natural ordering.

Although there are many other aspects of latent structure analysis which are discussed in Chapters 10 and 11—pertaining to the existence of more than one latent set of classes and to the characteristics of various types of questions—these cannot be touched upon in this review. Perhaps more important than those specific points, however, is the general comparison of scale analysis and of latent structure theory. The fact that the scale pattern appears as a special case of the latent structure theory does not at all detract from the usefulness of scale analysis. Provided that the hypothesis of scalability is satisfied, there appears to be more intuitive meaning which can be attached to the relationship between population and universe of content than will ever be possible in the more complex models. Moreover, the internal analyses possible for the scale pattern (following from the theory of principal components), have not as yet been generalized to the other types of response patterns. Finally, as will be apparent in the next section, the practical application of scale theory is much simpler than that of latent structure theory. It is naturally understood that both of these models may have to be changed considerably as experience develops. It is difficult to see how a model can remain static in a field as complex as that of attitude measurement.

5. *The Practical Application of Latent Structure Theory*

(a) *The choice and determination of a particular latent structure.* Even if one is given a population of individuals and a universe of questions which are known to satisfy the latent structure model, the problem still remains of actually spelling out the latent structure. In particular, it is vitally necessary to specify the number of classes which the latent attribute possesses. Provided that there exists such a perfect latent structure, Lazarsfeld has devised a criterion to tell how many of these classes there must be. Unfortunately, this criterion seems to be of theoretical interest only and of little practical importance. The practical difficulties are attributable to two principal sources. In the first place, perfect latent structures are no more likely to exist than are perfect scales; and in the second place, for any sizeable universe of questions, the computations would be wellnigh hopeless.

The foregoing points mean that a latent structure must be fitted to a population and universe of questions on a more or less "cut and try" basis. In other words, a value of λ (number of latent classes) is assumed and one proceeds to determine a "best fitting" structure for the manifest data. That is, "best fitting" values of n_I, n_{II}, $\cdots$, $n\lambda$, p_{I1}, $\cdots$, p_{Im}, p_{II1}, $\cdots$, $p_{\lambda m}$ are computed. From these latent structure parameters it is possible to predict the number of people who should fall in the various response patterns determined by the questions. Comparisons between the observed and fitted frequencies then gives a measure of fit. This, of course, immediately raises the problem of what constitutes a reasonable degree of fit. Although the problem is recognized and discussed, Volume IV offers very little guidance in this respect, and serious thought needs to be given in this phase of latent structure analysis. In particular, it seems that more precise statements concerning the uses which are to be made of the latent structure must be given before questions relating to "reasonableness of fit" can be answered. In the event that the classes of the latent attribute possess a natural ordering, the answer seems to be that a hypothesis of unidimensionality is being tested. If such natural ordering does not exist, the question still remains open.

It should be noted that a measure of fit derived in the above manner does not correspond to the coefficient of reproducibility used in scale analysis. Such a measure of fit is based upon a consideration of individuals whereas the coefficient of reproducibility is based upon a consideration of responses. In a crude sort of way, the corresponding value for scale analysis would be determined from the number of

non-perfect scale individuals.

In addition to the above mentioned considerations that are involved in arriving at a "best fitting" structure, there are two other points which can cause no end of difficulty. A tentative value of λ must be obtained and the structure parameters must be computed. At the present time, the determination of λ would seem to rest upon more or less intuitive and empirical grounds. Limited experience with War Department data has led to the development of certain rather general observations concerning the relationship between question form and question content and reasonable values of λ. Furthermore, certain structures are not computationally feasible. This means that attention must usually be focussed upon a restricted set of latent structures (e.g., the latent dichotomy, or the quasi-scale type of structure). Fortunately, those structures for which computations are possible are most likely to be those for which a natural ordering of the latent classes is provided. This may have a salutary effect in that effort will not be wasted, for a time at least, in deriving latent structures with which no particular meaning can be associated.

(b) *Testing the hypothesis of an approximate latent structure.* In order to apply latent structure theory to a specific situation, it is necessary to go through about the same steps as for the application of scale analysis. A population of individuals and a universe of questions are defined; a sample is selected from each and the responses (i.e., the manifest data) are obtained; a specific latent structure is postulated from past experience, knowledge of question form and content, and from consideration of what is computationally feasible; the fit of this structure to the date is evaluated; the results of the fitting procedure are used to infer something about the corresponding relationship in the entire population and universe; and finally, if the fit is too "bad," either a more complex structure if fitted or else the hypothesis of an approximate latent structure is rejected.

Most of these steps require no further discussion (e.g., the sampling of individuals), or the comments made in the preceding subsection will hold (e.g., assumption of a structure for which the computations can be carried out, necessity for having some form of natural ordering in order to provide meaning for the fitted structure, and so on). However, there is one extremely important point which does not seem to have received adequate attention, at least insofar as the material in Volume IV is concerned. This is the problem of question sampling. For example, suppose that a latent structure with three classes is fitted to the responses of the entire population on, say, four

questions (e.g., assume that n_I, n_{II}, n_{III}, p_{I1}, $\cdots$, p_{I4}, p_{II1}, $\cdots$, p_{II4}, p_{III1}, $\cdots$, p_{III4} are determined) and that the fit is "reasonable." If several new questions are added, say questions 5 and 6, and the same structure if fitted to all six questions (the four old ones and the two new ones), how will the new structure parameters compare with the old ones?

There is no reason to expect that the answers to questions of this type will be simple. As a matter of fact, the answers must be, by the very nature of the problem, extremely complex. However, it would seem that some progress might be made on a purely logical basis, to say nothing of the empirical studies that could be carried out. The need for this material is well illustrated by the following quotation taken from Chapter 1: "It should be said that certain of the problems of indeterminancy which haunt quantitative factor analysis appear in latent attribute analysis. Work with this new tool is still too young to have developed a completely standardized set of criteria for determining when the approximations are 'good enough' approximations. Such standards, doubtless, will be forthcoming but they wait upon a large amount of further empirical and theoretical investigation. For example, two items which seem by a priori inspection of content to be in the same general attitude area as the four items in our illustrative example were substituted for items 2 and 4. Although by conventional item analysis techniques, these new items should belong in the scale, they did not satisfy the criteria for the latent dichotomy model." The example referred to involved two latent classes and four dichotomous questions.

This discussion is not meant to imply that these points are not recognized in Chapters 10 and 11. They are recognized. However, one is left with the general impression that too much emphasis has been placed on the purely formal aspects of fitting a latent structure. This is not at all surprising and comes as a direct result of the recency of latent structure theory, the complexity and generality of the model, and the heavy analytic problems involved. More thought and research needs to be devoted in the near future to the problems of question-sampling and to the real meaning of the simpler forms of the model.

6. *Summary*

The principal points brought out in the course of this review are:

(1) Scale analysis and latent structure analysis provide models and objective procedures for approaching the problem of attitude

measurement. However, neither of them defines an attitude. Both deserve careful attention from a logical and from an empirical point of view. Such an examination may bring about changes in the models, but this is the only way in which progress can be made.

(2) Although the scale model can be viewed as a special case of the latent structure model, there would seem to be more intuitive meaning associated with a scale than with the more general case. Moreover, certain internal analyses can be made within the scale model which have not as yet been carried over to the other. The practical application of scale analysis is less troublesome than that of latent structure theory, but it may be applicable in a much narrower range of cases.

(3) Perhaps the single most important problem, and this is common to both models, is that of defining the universe of content (or of questions), sampling from this universe, and inferring from these results something about the entire universe of content.

(4) As far as the contents of Volume IV are concerned, latent structure theory suffers in contrast with scale analysis. In particular, less space is devoted to latent structure theory, a body of experience has not yet grown up from its application, considerable mathematical analysis and numerical computation are needed, and not enough thought has yet been given to the problems arising out of the latent structure model. It should be recognized that these facts are due substantially to the recency of the latent structure model, and they should not lead research workers to give latent structure analysis less attention than it deserves.

REFERENCES

1. Edwards, A. L., and Kilpatrick, F. P. A technique for the construction of attitude scales. *J. appl. Psychol.*, 1948, **32**, 374-384.
2. Festinger, L. The treatment of qualitative data by 'scale analysis.' *Psychol. Bull.*, 1947, **44**, 149-161.
3. Ford, R. A rapid scoring procedure for scaling attitude questions. *Publ. Opin. Quart.*, 1950, **14**, 507-532.
4. Guttman, L. The Cornell technique for scale and intensity analysis. *Educ. Psychol. Meas.*, 1947, **7**, 247-279.

Manuscript received 11/1/50
Revised manuscript received 2/10/51

PSYCHOMETRIKA—VOL. 16, NO. 3
SEPTEMBER, 1951

A SQUARE ROOT METHOD OF SELECTING A MINIMUM SET OF VARIABLES IN MULTIPLE REGRESSION: I. THE METHOD

A. SUMMERFIELD
UNIVERSITY COLLEGE, LONDON
AND
A. LUBIN
INSTITUTE OF PSYCHIATRY, MAUDSLEY HOSPITAL

An extension of Dwyer's "square root" method has been made to the problem of selecting a minimum set of variables in a multiple regression problem. The square root method of selection differs from the Wherry-Doolittle method primarily in that (1) the computations required are more compact, (2) an F ratio criterion is used which leads to the selection of fewer variables. The method provides solutions for the problems of test selection, item analysis, analysis of variance with disproportionate frequencies, and other problems requiring the rejection of superfluous variables. In a subsequent article a worked example will be given, and the square root and Wherry-Doolittle methods compared.

I. *Introduction.*

The purpose of the present paper is to introduce a compact procedure for selecting the minimum number of effective independent variables in a multiple regression problem. The problem has been previously discussed by Wherry (24), who developed a technique based on the Doolittle method. The procedure presented here is based on what Dwyer (4) has called the *square root* method of computing multiple correlations and regressions.

R. L. Thorndike (19, pp. 201-202), in addition to giving an admirable summary of the problem of selecting a battery of tests, has given a description of Wherry's method which is followed below. With some modifications the description applies to the square root selection method proposed here. Wherry's technique permits the successive addition of variables. Starting with the most valid single test, the second variable which will add most to the validity of the first is found. This involves determining the partial correlations of each test with the criterion when the first test is held constant. The regression weights and multiple correlation for the two tests are found, and a third test sought. The process is continued step by step; at each step the test that will give the greatest increase in the multiple corre-

lation is added.

Although the square root method follows this description on the whole, it differs from the Wherry-Doolittle method in several respects which we consider crucial.

(1) The multiple R^2 and beta weights for the whole battery of tests are computed as a first step. Whether any test selection at all is justified, can be determined by testing this multiple R^2 for significance. The beta weights serve as a partial check on the final test selection. In general, the selected tests can be expected to include all tests having significant beta weights, but other tests may also be selected.

(2) The decision procedure for ending the selection of additional tests depends on the use of the well known F ratio criterion (**16**, p. 266 and **23**, p. 262). F ratio tests are made to determine whether (a) the increase in the multiple R^2 introduced by an additional variable is greater than chance and (b) the multiple R^2 for the selected variables is significantly lower than the value obtained using all the tests in the battery. Wherry's decision procedure is shown to be too weak; it leads to the selection of more tests than are necessary.*

(3) The computational procedure is based on Dwyer's square root method rather than the Doolittle method. The computations are more compact, the coefficients easier to interpret, and the tabulations fewer, than for Wherry's method. The authors agree with Dwyer that the square root technique is an aid "toward fewer errors, less computing time and more pleasure for the computor" (**4**, p. 493). The method is particularly advantageous if automatic desk calculators are used.

II. *Relation of the square root method to factor analysis, partial and multiple correlation.*

The square root method is essentially similar to other regression methods in that it involves the reduction of the correlation matrix, R, of the n independent variables, to an $n \times n$ triangular matrix, T, a matrix with all its elements above the principal diagonal equal to zero. The distinctive feature of the method is the direct calculation of a triangular matrix which has the property

*Professor Burt, in his laboratory notes (1942), has suggested a method of test selection based on what he terms the "method of hierarchical subtraction," rather than on the "square root method." Use of the F ratio for testing the significance of the increase (2,a, above) is mentioned. An account of this technique has been published recently. (Burt, Cyril. The numerical solution of linear equations. *Brit. J. Psychol.*, statist. Sect., 1951, **4**, 31-54).

$$TT' = R . \tag{1}$$

Factor analysts will at once recognize expression (1) as a particular case of the fundamental theorem of factor analysis embodied in the more general expression, e.g., Thurstone (21, p. 66 and 22, p. 78),

$$FF' = R \tag{2}$$

where F is in general, an $n \times r$ matrix of factor saturations, r being the rank of R.

Thurstone (20, and 22, p. 101) has given a procedure, which he called the "diagonal method of factoring," for deriving a triangular (or trapezium) matrix from the correlation matrix. It is identical with Dwyer's procedure. Thurstone, of course, derives his triangular matrix from a correlation matrix having communalities in the principal diagonal, as he wishes to treat it as a matrix of common factors. The use of such a procedure in factor analysis was also suggested by Burt (1, footnote to p. 307).

The matrix T is therefore a triangular factor matrix obtained by the square root method from a correlation matrix with unities in the principal diagonal. We shall refer to it as the *square root matrix* in the context of this article. To call it a "diagonal" matrix would be misleading, as in matrix algebra this would mean a square matrix where "all the elements other than those in the principal diagonal are zero." (6, p. 3) And to use the term "triangular" matrix would not differentiate it from other triangular matrices derived from R, that would not satisfy equation (1). The term "square root" was first used to describe this triangular factor matrix by Dwyer (4). Holzinger and Harman (11, p. 94) in discussing the same procedure, state that 'This solution is obtainable by means of a general algebraic procedure for factoring any symmetric matrix, known as "completing the square." The method was applied specifically to a correlation matrix by McMahon [(15)] before 1923.' Holzinger (10) refers to it as the "solid staircase" method.

In interpreting a square root matrix, e.g., Table 1, as a factor matrix, the columns represent orthogonal factors, and the rows represent the projections of each test vector on the orthogonal axes. These projections are the correlations of each test with the orthogonal factors and therefore represent the regression coefficients for obtaining the test scores from the factor scores. Since each test vector has a length of unity, the factor saturations are actual correlations and the test scores can be obtained exactly from the factor scores.

The essential aspects of the square root method can be simply

TABLE 1
The Square Root Factor Matrix, T, in Terms of Semi-partial Correlations

VARIABLES	FACTORS						
	t_1 X_1	t_2 $X_{2.1}$	....	t_i $X_{i.1,2,\cdots,i-1}$	....	t_n $X_{n.1,2,\cdots,n-1}$	t_c $X_{c.1,2,\cdots,n}$
X_1	$r_{11}=1.00$	0	....	0	....	0	0
X_2	r_{21}	$r_{2(2.1)}$	....	0	....	0	0
....	....	....	....	....	....	....	....
X_i	r_{i1}	$r_{i(2.1)}$	....	$r_{i(i.1,2,\cdots,i-1)}$	....	0	0
....	....	....	....	....	....	....	....
X_n	r_{n1}	$r_{n(2.1)}$	....	$r_{n(i.1,2,\cdots,i-1)}$	....	$r_{n(n.1,2,\cdots,n-1)}$	0
X_c	r_{c1}	$r_{c(2.1)}$	....	$r_{c(i.1,2,\cdots,i-1)}$	....	$r_{c(n.1,2,\cdots,n-1)}$	$r_{c(c.1,2,\cdots,n)}$

described. A first axis, factor, or reference vector is identified with one of the variables. A second variable is taken and a second factor, orthogonal to the first, is derived to account for the residual variance of the second variable after the extraction of its communality in factor I. It follows that the first variable must have a zero loading on factor II. A third factor is made to account for the residual variance of a third variable, after the removal of its communality in factors I and II. Both first and second variables must now have zero correlations with factor III. A fourth variable is next taken and treated similarly, and so on.

It is very useful to think of the coefficients of the square root matrix as correlation coefficients. The first column in Table 1 represents the correlations between each of the tests and the first factor, which is simply the first test. The general element of this column is r_{i1} where i is any test.

The second column in Table 1 gives the correlations of each test with the 2nd factor, $X_{2.1}$. $X_{2.1}$ is the score on the second test when the linear effect of the first test is held constant. It is defined by the following equation:

$$X_{2.1} = X_2 - (a_{21} + b_{21}X_1) \tag{3}$$

where $a_{21} + b_{21}X_1$ is the least-squares prediction of X_2 based on X_1.

The general element of the second column may be called $r_{i(2.1)}$, the correlation of the ith test with $X_{2.1}$, that part of the X_2 score which can not be predicted by X_1. Professor Burt has suggested that these coefficients are equivalent to what are known as *semi-partial correlations,* and we shall refer to them as such; the reader should note that they are not the usual partial correlations.

The general element of the third column of Table 1 is $r_{i(3.1,2)}$, the correlation of the ith test with $X_{3.1,2}$, that part of each X_3 score that cannot be predicted from X_1 and X_2.

Obviously $r_{1(2.1)}$ must be zero since X_1 can only correlate zero with a score from which all linear influence of X_1 has been subtracted. Similarly,

$$r_{1(3.1,2)} = r_{2(3.1,2)} = 0. \tag{4}$$

The nth column of Table 1 consists entirely of zeros except for

$$r_{n(n.1,2,3,\cdots,n-1)} \quad \text{and} \quad r_{c(n.1,2,3,\ \cdots,n-1)}.$$

The last row of Table 1 is the set of correlations of the criterion test, c, with the n orthogonal factor scores. The first n columns represent a set of orthogonal scores which are capable of reproducing

the test intercorrelations exactly. The squared multiple correlation of the criterion, c, with the n orthogonal factor scores, $R^2_{c(n)}$, is equal to the sum of the n squared correlations in row c. Now, the n test scores can be reproduced exactly by the n factor scores. Therefore, the multiple R^2 of test c with the n test scores is exactly equal to $R^2_{c(n)}$, i.e., to the squared multiple correlation of the criterion with the n orthogonal factor scores. Moreover, since the first i tests are completely dependent on the first i factor scores, the squared multiple correlation of the criterion with the first i tests, $R^2_{c(i)}$, is the sum of squares of the first i correlations in row c. Evidently, $r^2_{c(i+1.1,2,3,\ldots,i)}$ is the additional contribution to this squared multiple correlation made by the $(i+1)$th test.

In general, the basic theorem used throughout this article is "Given the intercorrelations of n variables, the squared multiple correlation of the ith variable with k other tests, is equal to the communality of the ith variable on the k factors which reproduce the k other tests." This theorem holds because each test vector is taken to be of unit length, with no specifics.

General theorems of this type connecting multiple correlation with factor analysis, have been given by Roff (18), Dwyer (2, 3), and Guttman (8). Computing formulas, for both partial and multiple correlations have been given by Dwyer (3).

The equations for calculating the beta weights

Dwyer (4, pp. 497-499) points out that multiple regression weights for each of the variables can be calculated directly from a square root matrix such as Table 1. In essence, a "back solution" is made for the beta weights similar to the Doolittle procedure. We give a matrix algebra proof and then outline the technique in somewhat more detail than is given by Dwyer.

Let R be the $n \times n$ correlation matrix of the n independent variables. B is the $n \times 1$ column vector of beta weights whose general element of the jth row is β_j. Let r_c be the $n \times 1$ column vector of correlations of the criterion with the n independent variables. The element in the jth row of r_c will be denoted by r_{jc}.

The usual "normal" equations can be written;

$$RB = r_c. \tag{5}$$

From equation (1), $TT' = R$, so if we substitute for R we get

$$TT'B = r_c. \tag{6}$$

If equation (6) is premultiplied by T^{-1}, then

$$T'B = T^{-1}r_c. \tag{7}$$

The column vector $T^{-1}r_c$ represents the correlations of the criterion with the n orthogonal factors of T. This same set of correlations is shown as the row of correlations for the criterion variable in Table 1.

The column vectors of T in Table 1 are denoted by t, and t_j will be the jth column of T. The element of the ith row and jth column in T will be denoted by t_{ij}. Equation (7) implies that each of the column vectors of T when multiplied by the beta weights, will yield the semi-partial correlation of the criterion with the orthogonal factor represented by that column vector. In summation notation

$$B' t_j = \sum_{i=1}^{n} \beta_i t_{ij} = r_{c(j.1,2,\cdots,j-1)}. \tag{8}$$

Working backwards, we first solve for β_n, the last beta weight. We use t_n of Table 1, which has zeros everywhere except in the nth row and in the row for the criterion. Equation (8) when applied to t_n gives an equation with one unknown,

$$\beta_n r_{n(n.1,2,\cdots,n-1)} = r_{c(n.1,2,\cdots,n-1)}. \tag{9}$$

In terms of the elements of the t_n vector,

$$\beta_n t_{nn} = t_{cn}. \tag{10}$$

The solution for β_n is simply

$$\beta_n = t_{cn}/t_{nn} = \frac{r_{c(n.1,2,\cdots,n-1)}}{r_{n(n.1,2,\cdots,n-1)}}. \tag{11}$$

Having obtained β_n, the next weight, β_{n-1} can now be found by using equation (8) with t_{n-1}. There are only three non-zero entries in this column. Equation (8) is

$$\beta_n t_{n,n-1} + \beta_{n-1} r_{(n-1)(n-1.1,2,\cdots,n-2)} = r_{c(n-1.1,2,\cdots,n-2)}. \tag{12}$$

Since β_n is known numerically from equation (11) there is only one unknown, β_{n-1} and we can solve for it. In exactly the same fashion, all of the other multiple regression weights can be found.

It is recommended that the computation

$$R^2_{c.1,2,\cdots,n} = \sum_{i=1}^{n} \beta_i r_{ic} \tag{13}$$

should always be carried out. It is a useful, though not infallible check

on the computation of the beta weights.

III. *The F ratio criterion for determining when additional variables add significantly to the multiple correlation.*

Various methods of systematic test selection have been published; Dwyer (3), Horst and Smith (12), Johnson (13), Wherry (24): in none of these cases has an F ratio criterion been used to decide when to terminate the selection of variables.* We propose to employ the following well known test of significance.

Let R^2_n be the squared multiple correlation based on n variables. Let R^2_i be the squared multiple correlation based on a sub-set, i, of the n variables, $(i < n)$. To test the significance of the difference between R^2_n and R^2_i we take

$$F = \frac{(R^2_n - R^2_i)/(n-i)}{(1-R^2_n)/(N-n-1)}, \tag{14}$$

with degrees of freedom, $n_1 = n - i$ and $n_2 = N - n - 1$.

The foregoing paragraph is paraphrased from McNemar (16, p. 266). A general proof that this F ratio is a likelihood ratio criterion is given by H. B. Mann (17, Ch. IV).

In the case of the square root method, when $n = i + 1$, i.e., we wish to test the contribution of the $(i + 1)$th variable to R^2_i, then $(R^2_{i+1} - R^2_i)$ is equal to the squared correlation of the criterion with the $(i + 1)$th factor. (See Table 1.)

The F ratio is used here primarily as a criterion in a decision procedure, not as a test of significance giving exact probability levels. The difficulties of interpreting significance levels when F ratio tests are made sequentially are discussed in section V.

IV. *The square root method of selecting effective independent variables.*

The method of selecting variables that we are proposing follows at once from the properties the square root matrix has been shown to possess. The multiple R^2 of the criterion variable with the complete battery is determined. The hypothesis that this coefficient does not differ significantly from zero is tested to ensure that the battery possesses predictive power. The independent variable having the high-

*Professor Burt's method includes the use of this procedure. (See P. E. Vernon's *Notes on Statistical Methods*, 1945, pp. 6-11, where it is described in full with an illustrative example.

est validity is then selected, X_{i^*}, say. Its column of correlations in R is made the first column, t_1, of T. The multiple R^2 thus far is

$$R^2_{c.i^*} = r^2_{ci^*}. \tag{15}$$

As before, X_c is the criterion variable.

The semi-partial correlations of all $(n-1)$ remaining variables with X_c are now calculated. The set of quantities, $r_{c(j.i^*)}$, $(j = 1, 2, \cdots, i^*-1, i^*+1, \cdots, n; j \neq i^*)$, is thus obtained. As has been shown, the squares of these quantities are the contributions to the multiple R^2 that each of the remaining independent variables can make individually, the variable X_{i^*} having been selected. The highest value of $r_{c(j.i^*)}$, given by variable X_{j^*} say, is taken. Hence

$$R^2_{c.i^*,j^*} = r^2_{ci^*} + r^2_{c(j^*.i^*)}. \tag{16}$$

The difference between (15) and (16) is tested by (*vide* section III),

$$F = \frac{(r^2_{c(j^*.i^*)})/1}{(1-R^2_{c.i^*,j^*})/(N-2)}, \quad [\text{d.f.}: n_1 = 1, n_2 = N-2]. \tag{17}$$

If the F ratio is found to be significant, X_{j^*} becomes the second variable selected. Its column in the $(n+1)\times(n+1)$ correlation matrix is used to compute the second column, t_2 of T.

The semi-partial correlations, $r_{c(k.i^*,j^*)}$, of the remaining $n-2$ independent variables, $(k = 1, 2, \cdots, i^*-1, i^*+1, \cdots, j^*-1, j^*+1, \cdots, n; k \neq i^*, j^*)$, are next calculated. The highest value, that for variable X_{k^*}, say, is taken, giving

$$R^2_{c.i^*,j^*,k^*} = r^2_{ci^*} + r^2_{c(j^*.i^*)} + r^2_{c(k^*.i^*,j^*)}. \tag{18}$$

The difference between (16) and (18) is tested, as before, and if the increment due to X_{k^*} is found significant, X_{k^*} becomes the third variable selected, t_3 is calculated; and so on.

When a stage is reached at which none of the remaining independent variables makes a further significant contribution to the multiple R^2, the process terminates. It remains possible that if one or more of the variables not so far selected exerts a suppressor effect, then some pair or higher-order combination of the remaining variables might together add significantly to the multiple R^2. An advantage of the method is that it permits a test to be made against this contingency. Suppose that the selection process has been terminated according to the criterion just given, when p variables have in fact been selected which jointly yield a squared multiple correlation of

$R^2_{c.(p)}$. The multiple correlation of the criterion variable, X_c, with the whole battery has already been calculated (*vide* first paragraph of this section). If it is shown that $R^2_{c.(p)}$ does not differ significantly from $R^2_{c.1,2,\cdots,n}$ then it follows that none of the $n - p$ remaining variables, neither singly nor jointly in any combination, can add any further significant increment to $R^2_{c.(p)}$. The test is:

$$F = \frac{(R^2_{c.1,2,\cdots,n} - R^2_{c.(p)})/(n-p)}{(1 - R^2_{c.1,2,\cdots,n})/(N-n-1)}, \tag{19}$$

with degrees of freedom, $n_1 = n - p$ and $n_2 = N - n - 1$. Should the test indicate a significant difference, it would be necessary to search for its source; otherwise it would be concluded that the p variables selected were an optimum set of effective independent variables. The selection process should stop only when the multiple R^2 based on the selected variables does not differ significantly from the R^2 based on the whole battery. In actual applications of the method for test selection it has not so far been found necessary to select more than four variables. For example, in a student selection study by Himmelweit and Summerfield (9) out of a battery of twenty-one tests only four were selected.

V. *The merits and demerits of the square root method of test selection.*

The selected tests will generally have the highest beta weights. The beta weight, roughly speaking, measures the contribution of a variable to the multiple correlation, independently of all other predictor variables. Therefore, those variables with large beta weights in the complete battery are most likely to be selected for the reduced battery. But variables whose beta weights do not differ significantly from zero may, when selected, make a significant contribution. Their near-zero regression coefficient may arise from the fact that some other variables measure very nearly the same thing. Fisher (5, p. 422) cites a study where

> . . . when seven sea-level characteristics were employed in the prediction, not one of the coefficients was significant, although an apparently good prediction was obtained from the multiple regression formula. All that the non-significance meant, however, was that if any one of the coefficients were given the value zero *and the other coefficients adjusted*, the prediction formula was not significantly impaired. The sea-level characteristics showed, in fact, sufficiently close mutual correlation for any one of them to be capable of replacement by an appropriate linear func-

tion of the others, so as to compensate nearly completely for its absence from the prediction formula.

The *square root method of selection* is less likely than the Wherry-Doolittle method to terminate with variables having non-significant beta weights. Of course, any method which selected the variables making the highest contribution to the multiple correlation and which used the F ratio tests that we advocate would have the same result. There is some possibility that one of the earlier variables selected will be highly correlated with later variables, but we suggest that this would be quite rare. The method would therefore seem to be a solution to Ragnar Frisch's problem of "nonsense multiple correlations" (7). But even if the square root method were always to give significant beta weights, it does not follow that the "best" solution has been found.

The "best" solution could be found if we were to proceed as follows:

(1) calculate $R^2_{(n)}$ based on all n variables;

(2) calculate $R^2_{(1)}$ for each single variable and test each of these coefficients for significance against $R^2_{(n)}$;

(3) calculate $R^2_{(2)}$ for each of the $\frac{n(n-1)}{2!}$ pairs of variables and test each of the coefficients for significance against $R^2_{(n)}$;

(4) calculate $R^2_{(3)}$ for each of the $\frac{n(n-1)(n-2)}{3!}$ triads of variables and test each coefficient for significance against $R^2_{(n)}$; and so on, until a value of $R^2_{(q)}$ not significantly different from $R^2_{(n)}$ was obtained. If, then, the null hypothesis were accepted for, let us say, a particular group of q variables, then this group would necessarily be the "best" set of variables to take. For all other sets of q or fewer variables would have been tested and found wanting.

Now it should be clear that although the "best" set of variables, q, may involve less than the number of variables, say p, selected by the square root method, p will never be less than q. This must be the case since the square root method ensures that the $R^2_{(p)}$, based on the p selected variables, can never differ significantly from $R^2_{(n)}$.

The square root method of selection is a special case of what Dr. M. P. Schützenberger has called a "one-step locally best" solution to a maximisation demanding a finite sequential procedure. Dr. Schützenberger (personal communication) has given a general theorem to the effect that when the steps are independent of one another, the "best" solution is the "one-step locally best" solution. In our case, he says, this means that when all tests have zero intercorrelations, the

square root method gives the best possible solution, but it may also give the best possible solution in other cases as well.

The selection method discussed here can be extended to such problems as item analysis and analysis of variance with disproportionate frequencies, to reject superfluous explanatory variables. In connection with the analysis of variance problem, James Durbin (personal communication) has pointed out that:

> . . . the procedure of selecting the explanatory variable which gives the largest value of F alters the significance level of the test.
>
> Suppose for simplicity that the null hypothesis is known to be true. Then if the p values of F were independent of each other, the probability that none would be significant at the 5% level would be
>
> $$\left(1-\frac{1}{20}\right)^p.$$
>
> Thus the probability that at least one value appeared significant would be
>
> $$1-\left(1-\frac{1}{20}\right)^p.$$
>
> This is greater than 1/20 if $p > 1$. In fact the values of F would not be independent but would be correlated, though not perfectly. Thus the effective significance level used would be somewhere between 1/20 and
>
> $$1-\left(1-\frac{1}{20}\right)^p.$$
>
> In the non-null case the complications are likely to be even greater.

(See also Kendall on "Use of the z test for several variance ratios," 14, p. 199 ff.)

The foregoing criticism applies equally to the present case of multiple regression. Nevertheless, we suggest that the F ratio test provides the best available statistical decision procedure. However, the tabled levels of significance of F evidently do not apply precisely. It is clear that the square root method tends to select rather too few tests whereas the Wherry-Doolittle leads to the selection of too many.

Another defect of the square root method (and all other selection procedures) is that the method capitalises sampling errors. It is recommended that the population estimate of the squared multiple correlation be calculated as the "shrunken" $R^2_{(n)}$, where all n variables are used in calculating the shrinkage. This is, of course, the estimate against which we have compared multiple correlations based on selected variables.

Acknowledgments

We would like to acknowledge the various criticisms our manu-

script has received from Professors Sir Cyril Burt and Sir Godfrey Thomson. We are also indebted to our colleagues, especially J. Tizard, N. O'Connor, and Dr. A. Clarke, who have used the method and helped us overcome some practical difficulties. Our thanks are due to Miss D. Webb and Miss A. Davis for secretarial assistance.

REFERENCES

1. Burt, Sir Cyril. The factors of the mind. London: Univ. of London Press, 1940.
2. Dwyer, P. S. The contribution of an orthogonal multiple factor solution to multiple correlation. *Psychometrika*, 1939, 4, 163-171.
3. Dwyer, P. S. The evaluation of multiple and partial correlation coefficients from the factorial matrix. *Psychometrika*, 1940, 5, 211-232.
4. Dwyer, P. S. The square root method and its use in correlation and regression. *J. Amer. stat. Ass.*, 1945, 40, 493-503.
5. Fisher, R. A. The precision of discriminant functions. *Ann. Eugen.*, 1940, 10, 422-429.
6. Frazer, R. A., Duncan, W. J., and Collar, A. R. Elementary matrices. Cambridge: Cambridge University Press, 1947.
7. Frisch, R. Statistical confluence analysis by means of complete regression equations. Pub. No. 5. Universitets Okonomiske Institut, Oslo, 1934.
8. Guttman, L. Multiple rectilinear prediction and the resolution into components. *Psychometrika*, 1940, 5, 75-99.
9. Himmelweit, H. T., and Summerfield, A. Student selection: II. *Brit. J. Sociol.*, 1951, 2, 59-75.
10. Holzinger, K. J. Preliminary report on Spearman-Holzinger unitary trait study, No. 5. Chicago: Statistical Laboratory, Department of Education, Univ. of Chicago, 1934.
11. Holzinger, K. J., and Harman, H. H. Factor analysis. Chicago: Univ. of Chicago Press, 1941.
12. Horst, P., and Smith, S. The discrimination of two racial samples. *Psychometrika*, 1950, 15, 271-290.
13. Johnson, P. O. Statistical methods in research. New York: Prentice-Hall, 1949.
14. Kendall, M. G. The advanced theory of statistics, Vol. II. London: Charles Griffin, 1948.
15. McMahon, J. Hyperspherical goniometry; and its application to correlation theory for n variables. *Biometrika*, 1923, 15, 173-208.
16. McNemar, Q. Psychological statistics. New York: Wiley. London: Chapman and Hall, 1949.
17. Mann, H. B. Analysis and design of experiments. New York: Dover Publications, 1949.
18. Roff, M. Some properties of the communality in multiple factor theory. *Psychometrika*, 1935, 1, 1-6.
19. Thorndike, R. L. Personnel selection: test and measurement techniques. New York: Wiley, 1949.
20. Thurstone, L. L. Theory of multiple factors. Ann Arbor, Michigan: Edwards Bros., 1932.

21. Thurstone, L. L. The vectors of mind. Chicago: University of Chicago Press, 1935.
22. Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. of Chicago Press, 1947.
23. Tippett, L. H. C. The methods of statistics. London: Williams and Norgate, 1948 (3rd ed.).
24. Wherry, R. J. In Stead, W. H., and Shartle, C. P. Occupational counseling techniques. Appendix V. New York: American Book Company, 1940, pp. 245-252.

Manuscript received 2/5/51
Revised manuscript received 3/25/51

PSYCHOMETRIKA—VOL. 16, NO. 3
SEPTEMBER, 1951

EFFECT OF GROUP HETEROGENEITY ON ITEM PARAMETERS*

HAROLD GULLIKSEN
EDUCATIONAL TESTING SERVICE

Most indexes of item validity and difficulty vary systematically with changes in the mean and variance of the group. Formulas are presented showing how certain item parameters will vary with these alterations in group mean and variance. Item parameters are also suggested which should remain invariant under such changes. These parameters are developed under two different assumptions: first, the assumption that the *total distribution* of the item ability variable is normal, and, second, that the distribution of the item ability variable *for each array* of the explicit selection variable is normal.

Most indexes of item validity or item difficulty vary systematically with the ability level of the group and with the variance in the ability of the group. Thus, the "per cent of persons answering an item correctly" increases as the mean ability of the group increases and decreases as the mean ability of the group decreases. The "item-criterion correlation" increases and decreases as the variance of the group increases and decreases. Item parameters which did not vary systematically as the mean and variance of the group ability changed would be valuable. Lacking such parameters it would be valuable to have formulas indicating the amount of change in a given item parameter to be expected for a given change in group mean and variance.

This paper will present the derivation of formulas which show the amount of change in various item parameters which should occur because of changes in group heterogeneity and also will develop item parameters which do not systematically increase or decrease as the heterogeneity of the group changes.

We will consider two groups of persons. An *unselected* group with a specified mean and variance and a *selected* group with a different mean (usually though not necessarily higher) and a different variance (usually though not necessarily smaller). It should also be noted that the theory to be developed applies regardless of the direction of the change in mean and variance. Thus, rather than utilizing

*The writer wishes to acknowledge helpful discussions of this paper with Paul Horst and Herbert S. Sichel who have worked on various aspects of the problem of invariant item parameters.

one set of symbols to designate the selected group and another set to designate the unselected group we will say that lower-case letters will be used to designate the group *for which all the parameters are available*; upper-case letters will be used to designate the group for which some information is available and for which *information is desired on other parameters.* That is to say, the unknowns to be solved for, will always be designated by upper-case letters. It will be noted during the derivation that no assumption is made regarding the direction of the change so that the formulas hold for utilizing information from an unselected group to estimate parameters in a selected group or for utilizing parameters from a selected group to estimate those in an unselected group.

I. *Definitions and Assumptions*

We will consider three major types of variables, as follows:

X, x is used to designate the *explicit* selection variable, that is, the variable which was directly used for selection.

Y, y is used to represent any other variable on which selection occurs only because of its correlation with X (x); this variable may be termed the variable subject to *incidental* selection.

I_G, i_g designates a gross score variable which represents the ability required to answer item G (or g). ($G = 1 \cdots K$; $g = 1 \cdots k$.) For example, for a test of K (k) items there would be K (k) such variables, a different one for each item.

It is also necessary to assume that for each item ability variable there exists an *ability level* such that all persons above this level answer the item correctly and all persons below this level answer the item incorrectly.

I'_G, i'_g will be used to designate the particular ability score above which all persons answer item G (g) correctly and below which all persons answer item G (g) incorrectly.

For any given item it is assumed that

$$I'_G = i'_g \qquad (G = g). \tag{1}$$

It is also necessary to designate means, standard deviations, correlations, and proportions for the different variables.

M, m (with appropriate subscripts—X, x, Y, y, G, g) designates the means of the various groups.

S, s (again with appropriate subscripts—X, x, Y, y, G, g) designates the standard deviations.

R, r (with appropriate subscripts) designates correlations.

P_G, p_g designates the proportion of persons answering the item correctly.

Z_G, z_g designates standard scores for the variables I_G, and i_g respectively.

That is to say,

$$Z_G = \frac{I_G - M_G}{S_G}, \qquad z_g = \frac{i_g - m_g}{s_g}.$$

Z'_G, z'_g are standard score levels corresponding to P_G and p_g and to I'_G and i'_g; that is to say

$$Z'_G = \frac{I'_G - M_G}{S_G}, \qquad z'_g = \frac{i'_g - m_g}{s_g}. \tag{2}$$

From equations (1) and (2) we see that

$$Z'_G S_G + M_G = z'_g s_g + m_g. \tag{3}$$

In developing the theory for the influence of group selection on item parameters we will utilize the usual assumptions for the influence of selection on correlation (references 1, 2, 3, 4, 5, and 7). First, assume that the regression line of I_G on X is identical with the regression of i_g on x. To have identical regression lines we need to assume that both the *slope* and *intercept* of the regression line is the same for the two groups, e.g.,

$$R_{XG}\frac{S_G}{S_X} = r_{xg}\frac{s_g}{s_x} \qquad \text{and} \tag{4}$$

$$M_G - R_{XG}\frac{S_G}{S_X}M_X = m_g - r_{xg}\frac{s_g}{s_x}m_x. \tag{5}$$

It is also necessary to assume that the variance about each of these regression lines is the same for both groups. Thus, we have

$$S_G^2 - S_G^2 R_{XG}^2 = s_g^2 - s_g^2 r_{xg}^2. \tag{6}$$

Correspondingly, we assume that the regression of Y on X is the same as the regression of y on x. Thus, we have

$$R_{YX}\frac{S_Y}{S_X} = r_{yx}\frac{s_y}{s_x}, \qquad \text{and} \tag{7}$$

$$M_Y - R_{YX}\frac{S_Y}{S_X}M_X = m_y - r_{yx}\frac{s_y}{s_x}m_x. \tag{8}$$

It is also assumed that the variance about each of these regression lines is the same. Thus, we have

$$S_Y^2 - S_Y^2 R_{YX}^2 = s_y^2 - s_y^2 r_{yx}^2. \tag{9}$$

In addition to the foregoing six assumptions it is also necessary to assume that selection is such that the correlation between the item ability variable and the incidental selection variable with the explicit selection variable partialled out is the same for both groups. That is to say, it is assumed that $R_{YG.X} = r_{yg.x}$. Using the formula for partial correlation, we have

$$\frac{R_{YG} - R_{XG}R_{YX}}{\sqrt{1-R_{XG}^2}\ \sqrt{1-R_{YX}^2}} = \frac{r_{yg} - r_{xg}r_{yx}}{\sqrt{1-r_{xg}^2}\ \sqrt{1-r_{yx}^2}}. \tag{10}$$

II. *Effect of Selection on Mean and Variance*

We will first show the general formulas for the effect of selection on mean and variance. From equations (4) and (6) we see that

$$\frac{S_G^2}{s_g^2} - 1 = r_{xg}^2\left(\frac{S_X^2}{s_x^2} - 1\right). \tag{11}$$

The relation between the variance of the incidental and explicit selection variables can be obtained from equations (7) and (9) giving

$$\frac{S_Y^2}{s_y^2} - 1 = r_{yx}^2\left(\frac{S_X^2}{s_x^2} - 1\right). \tag{12}$$

In some cases when dealing with items the correlation r_{yx} will vary for each different selection of cases on each item while it could con-

ceivably be true that R_{YX} would be the same for all items since we are thinking of an unselected total group. In this case we may utilize formulas (7) and (9) to write

$$\frac{s_{..}^2}{S_Y^2} - 1 = R_{YX}^2 \left(\frac{s_x^2}{S_X^2} - 1 \right). \tag{13}$$

Either equation (12) or (13) shows the change in variance of the incidental and explicit selection variables as a function either of r_{yx} or R_{YX}. The formulas showing change in means are also needed. From equations (4) and (5) we have

$$m_g - M_G = r_{xg} \frac{s_g}{s_x} (m_x - M_X). \tag{14}$$

From equations (7) and (8) we have

$$m_y - M_Y = r_{yx} \frac{s_y}{s_x} (m_x - M_X), \quad \text{and} \tag{15}$$

$$m_y - M_Y = R_{YX} \frac{S_Y}{S_X} (m_x - M_X). \tag{16}$$

Equations (15) and (16) show the change in mean of the explicit and incidental selection variables as a function of r_{yx} or R_{YX}, whichever is the more convenient.

In the subsequent derivations we will assume that at least three of the four means (m_y, M_Y, m_x, M_X) are known. If any three are known, equations (15) or (16) may be utilized to obtain the other one. If all four are known, equations (15) and (16) may be utilized to check on the agreement of the data with the basic assumptions indicated by equations (7) and (8). Correspondingly, we assume that three out of the four standard deviations (s_y, S_Y, s_x, S_X) are known. If any three are known, equations (12) or (13) may be utilized to obtain the other one. If all four are known, equations (12) and (13) may be utilized to check the reasonableness of assumptions (7) and (9). It is also necessary to know either r_{yx} or R_{YX}. In addition it will be found that r_{xg}, r_{yg}, and z'_g must be known in order to utilize the formulas that are to follow.

Given the information indicated above we now turn to the problem of estimating the effect of selection on item difficulty and item validity indexes.

So far, the writer has not found it possible to complete the solu-

tion without some assumption regarding the distribution of the item ability variable (I_G, i_g). Two different assumptions regarding the distribution of this variable are presented. Section III develops a set of formulas under the assumption that variable i_g may be assumed to be normally distributed. Thus it becomes reasonable to estimate r_{xg} and r_{yg} by means of biserial correlation coefficients. Section IV utilizes the assumption as suggested by Gillman and Goode (3) that variable i_g is normally distributed about the regression of i_g on x for *each of the* x-*arrays*. The first assumption is that the marginal totals of i_g are normally distributed. The second assumption is that the *deviations* from the regression line for each array are normally distributed. It may be that there are some conditions under which one assumption is superior and some conditions under which the other assumption is superior. No studies have been made on this topic, so far as the writer is aware.

III. *A Normal Distribution Assumed for Item Ability* (i_g)

(a) Effect of Selection on Item Difficulty and Correlation.

If it is assumed that the distribution of i_g is nearly normal, we may obtain numerical values for r_{xg} and r_{yg} by use of the formula for the biserial correlation coefficient. It should be noted that the use of biserial correlation assumes only a normal distribution of i_g. No restrictions are implied by the biserial correlation coefficient on the variance of i_g. It should be noted also that p_g is known from the data and since i_g is assumed to be normally distributed, z'_g can be found as a function of p_g from tables of the normal curve. Thus we have, given by the data, numerical values for r_{xg}, r_{yg}, and z'_g provided we are willing to assume a normal distribution of i_g.

With this information, values of R_{XG}, Z'_G, and R_{YG} can be obtained by the following procedures. From equations (4) and (11) we have

$$R_{XG} = \frac{r_{xg} S_X}{\sqrt{r_{xg}^2 S_X^2 + s_x^2 (1 - r_{xg}^2)}}. \tag{17}$$

From equations (3), (11), and (14) we find

$$Z'_G = \frac{s_x z'_g + (m_x - M_X) r_{xg}}{\sqrt{r_{xg}^2 S_X^2 + s_x^2 (1 - r_{xg}^2)}}. \tag{18}$$

This equation gives Z'_G as a function of the means and standard devi-

ations of the explicit selection variable (m_x, M_X, s_x, S_X). It may be noted that Z'_G may also be expressed as a function of the means and standard deviations of the incidental selection variable (m_y, M_Y, s_y, S_Y) as well as the other variables (z'_g, r_{yx}, r_{xg}) by utilizing (12) and (15) in equation (18).

It will be noted that Z'_G is a standard score as defined by equation (2) representing the ability level on variable I_G required for answering item G. If one assumes a normal distribution for the variable I_G then P_G (the proportion of persons in the new group who would be expected to answer the item correctly) may be computed. In particular, if a normal distribution of I_G is assumed then tables of the normal curve may be utilized to obtain P_G from Z'_G.

By substituting values from equations (4), (6), (7), (9), (11), and (12) in (10) we find

$$R_{YG} = \frac{r_{yg}\, r_{yx} + r_{xg}\left(\frac{S_Y^2}{s_y^2} - 1\right)}{\frac{S_Y}{s_y}\sqrt{r_{yx}^2 + r_{xg}^2\left(\frac{S_Y^2}{s_y^2} - 1\right)}}. \tag{19}$$

It should be noted that (19) gives R_{YG} as a function of r_{yx}. If it is more convenient to utilize the value R_{YX} we obtain the appropriate equation by substituting (9) in (19) obtaining

$$R_{YG} = \frac{r_{xg}\left(\frac{S_Y^2}{s_y^2} - 1\right) + r_{yg}\sqrt{\frac{S_Y^2}{s_y^2}(R_{YX}^2 - 1) + 1}}{\frac{S_Y}{s_y}\sqrt{r_{xg}^2\left(\frac{S_Y^2}{s_y^2} - 1\right) + \frac{S_Y^2}{s_y^2}(R_{YX}^2 - 1) + 1}}. \tag{20}$$

Equations (19) and (20) require the value of both variances for the incidental selection variable, e.g., S_Y (s_y).

If it is more convenient to obtain R_{YG} in terms of the variances of the explicit selection variable (S_X, s_x) then the proper formula may be found by substituting (12) in (19) obtaining

$$R_{YG} = \frac{r_{yg} + r_{xg}\, r_{yx}\left(\frac{S_X^2}{s_x^2} - 1\right)}{\sqrt{1 + r_{yx}^2\left(\frac{S_X^2}{s_x^2} - 1\right)}\sqrt{1 + r_{xg}^2\left(\frac{S_X^2}{s_x^2} - 1\right)}}. \tag{21}$$

which gives R_{YG} in terms of r_{yx} and the variances S_X^2 and s_x^2. We may substitute equation (13) in (20) and simplify obtaining

$$R_{YG} = \frac{r_{xg} R_{YX}(S_X^2 - s_x^2) + r_{yg}\, s_x \sqrt{S_X^2 - R_{YX}^2(S_X^2 - s_x^2)}}{S_X \sqrt{r_{xg}^2(S_X^2 - s_x^2) + s_x^2}}, \quad (22)$$

which gives R_{YG} in terms of R_{YX} and the variances S^2_X and s^2_x.

Equations (17) to (22) inclusive show the changes to be expected in the item difficulty parameter and item correlation dependent on various values of mean and standard deviation of the group.

(b) Item Indexes Which Do Not Vary Systematically with Changes in Group Ability.

It is possible to devise item indexes that would not be expected to change systematically with group selection by the following procedure. Divide equation (4) by the square root of equation (6) and designate the ratio by A''_{gx}. This gives

$$A''_{gx} = \frac{r_{xg}}{s_x \sqrt{1 - r_{xg}^2}} = \frac{R_{XG}}{S_X \sqrt{1 - R_{XG}^2}}. \quad (23)$$

It will be noted that both equations (4) and (6) are assumed invariant with group selection. The unknown values s_g and S_G disappeared on dividing, so that equation (23) gives an invariant function of the standard deviation and correlation. The item index A''_{gx} may be converted into R_{XG} for any arbitrary standard deviation S_X by equation

$$R_{XG} = \frac{A''_{gx} S_X}{\sqrt{1 + A''^2_{gx} S_X^2}}. \quad (24)$$

Thus we see that if the item ability may be assumed to be nearly normally distributed A''_{gx} may be an invariant index from which the correlation R_{XG} could be obtained for groups with any specified variability (S_X).

With respect to item difficulty we may use X''_G to designate the ability level for variable X such that half the persons pass and half fail item G. This means that we must find the point where the regression of I_G on X crosses the critical ability level Z'_G. From equations (3), (4), and (5) we have

$$X''_G = \frac{z'_g s_x}{r_{xg}} + m_x = \frac{Z'_G S_X}{R_{XG}} + M_X. \quad (25)$$

This equation gives an invariant relationship that should represent

the ability level at which half the persons will pass and half fail item G.

As yet no useful invariant relationship between two *incidental* selection variables has been devised. Some invariant function of r_{yg} and R_{YG} would be rather interesting to obtain. For example, if equation (10) is multiplied by the square root of equations (6) and (9), divided by equation (4), and multiplied by equation (7), we see that

$$s_y^2 r_{yz}^2 \left(\frac{r_{yg}}{r_{yz} r_{zg}} - 1 \right) = S_Y^2 R_{YX}^2 \left(\frac{R_{YG}}{R_{YX} R_{XG}} - 1 \right). \qquad (26)$$

However, three different correlation coefficients are involved in this relationship. An index such as that indicated by equation (26) does not seem to be particularly useful at present.

IV. *Normal Distribution of Item Ability for Each* X-*Array*

Instead of assuming a normal distribution for the marginal totals on variable I_G we may assume that the distribution of I_G is *normal* (or nearly normal) *for each of the* X-*arrays.* This assumption has been presented by Gillman and Goode (3) and the index suggested by them utilized by Sichel (6). If we say that there are h different arrays of X we may use the subscript ($f = 1 \cdots h$) to designate the array and say that I_{fg} is assumed to be normally distributed for each value of f and g.

Let x_f be the mid-point of the fth array of x,

p_{fg} be the per cent of persons *in this* fth *array* who pass item g (e.g., the per cent above ability level i'_g for item g),

u_{fg} be defined as a base line deviation on a normal curve corresponding to the area p_{fg}. Thus for each array f, u_{fg} is given as a function of the known value p_{fg}.

We may also adopt the convention that

if $p_{fg} > 50$, then $u_{fg} > 0$;
if $p_{fg} < 50$, then $u_{fg} < 0$.

Thus, for any given item g we have pairs of values u_{fg} and x_f ($f = 1 \cdots h$). A plot of these values (u vs. x) can be made for each item. This method of analysis has the advantage that if the assumptions involved are correct the plot will be approximately linear or, conversely, we may say if the plot of u_{fg} against x_f ($f = 1 \cdots h$) for a given item g is markedly nonlinear it shows that the assumptions

made do not hold for that particular item. One thus has a check on the reasonableness of the basic assumptions. If the variables x and i_g (or its transform u_g) are linearly related then we have equation

$$u_{fg} = A_{gx}x_f - B_{gx}. \tag{27}$$

The values of A_{gx} and B_{gx} may be determined from the graph by some appropriate curve-fitting method. One may use any method from rough graphic approximations to a least-squares method. However, given the plot of u_{fg} against x_f for a given item, numerical values for A_{gx} and B_{gx} can be determined.

To interpret A_{gx} and B_{gx}, we note from the regression of i_g on x that the point on the regression line (which may be designated $\hat{i}_g - m_g$) is given by the expression $\hat{i}_g = i'_g + u_{fg}\, s_g \sqrt{1 - r_{xg}^2}$. Since this is the point on the regression line corresponding to any particular value such as x_f we may write

$$i'_g + u_{fg}\, s_g \sqrt{1 - r_{xg}^2} = r_{xg}\frac{s_g}{s_x}(x_f - m_x) + m_g. \tag{28}$$

If equation (28) is solved explicitly for u_{fg} as a function of x_f we find the following values for the constants A_{gx} and B_{gx} of equation (27)

$$A_{gx} = \frac{r_{xg}\, s_g}{s_x\, s_g \sqrt{1 - r_{xg}^2}} = \frac{r_{xg}}{\sqrt{1 - r_{xg}^2}}\left(\frac{1}{s_x}\right) \tag{29}$$

and

$$B_{gx} = \frac{r_{xg}\, m_x}{s_x \sqrt{1 - r_{xg}^2}} + \frac{i'_g - m_g}{s_g \sqrt{1 - r_{xg}^2}}. \tag{30}$$

Equations (29) and (30) thus give the interpretation of the empirically obtained constants A_{gx} and B_{gx}. It will be noted that A_{gx} as interpreted by equation (29) is identical with A''_{gx}, as defined in equation (23), and should have identical properties as an invariant item index.

We now turn to the problem of a suitable item difficulty index. At the point where $u_{fg} = 0$ we may specify that x_f is indicated by X_G. This point is the one where half the persons pass and half fail the item. If we set $u_{fg} = 0$ and $x_f = X_G$ in equation (27) and utilize equations (29) and (30) to solve for X_G we find

$$X_G = \frac{B_{gx}}{A_{gx}} = m_x + \frac{i'_g - m_x}{s_g} \frac{s_x}{r_{xg}}. \tag{31}$$

Utilizing (2), (31) may be rewritten as

$$X_G = m_x + z'_g \left(\frac{s_x}{r_{xg}} \right). \tag{32}$$

It may be noted that equation (32) is identical with equation (25). That is, the index X_G given by equation (31), is identical with the index X''_G given in equation (25).

Thus we see that if the assumptions used in the derivation are satisfied, A_{gx} (or A''_{gx}) and X_G (or X''_G) are item indexes which should be invariant with respect to changes in the standard deviation of the variable x. Equations (23) and (25) give values for these indexes derived from the assumption that the marginal totals of i_g for persons answering the item may be regarded as normally distributed. Equations (27), (29), and (31) give values for these indexes based on the assumption that the distribution of i_g for each array of x may be regarded as normal. An investigation of the constancy of these indexes as arrived at by the two different sets of formulas would indicate which assumption was the better.

It should be noted that by utilizing the approach indicated by Gillman and Goode (3) it would also be possible to obtain the correlation between y and i_g for each array of variable x. This correlation would correspond to the partial correlation between y and i—holding x constant. One would thus empirically verify the assumption of equal partial correlation which is utilized in equation (10). One might then use some invariant index such as that given in equation (26) to obtain R_{YG} for any distribution for which S_Y, S_X, and R_{YX} were known. The value of R_{XG} for use in equation (26) could be obtained by the use of equation (24).

V. *Summary*

In dealing with the problem of item indexes and group variability two different assumptions have been suggested. The assumption that the item ability variable is normally distributed gives equations (17) to (22) showing changes in item parameters with changes in group mean and variance. Under this assumption equations (23) and (25) give item indexes which should be relatively invariant with respect to changes in group mean and standard deviation. The assumption that there is a normal distribution for each x-array as uti-

lized by Gillman and Goode has also been suggested. By means of a curve-fitting procedure we may find values for the item parameters indicated in equations (27), (29), and (31). These parameters should be invariant with respect to changes in group mean and variance. An investigation of the conditions under which either of the sets of indices proposed would be relatively invariant with selection is still to be made.

REFERENCES

1. Aitken, A. C. Note on selection from a multivariate normal population. *Proc. Edinb. math. Soc.*, 1934, 4, 106-110.
2. Burt, Cyril. Statistical problems in the evaluation of Army tests. *Psychometrika*, 1944, 9, 219-235.
3. Gillman, Leonard, and Goode, Harry H. An estimate of the correlation coefficient of a bivariate normal population when X is truncated and Y is dichotomized. *Har. educ. Rev.*, 1946, 16, 52-55.
4. Kelley, T. L. Statistical Methods. New York: The Macmillan Company, 1923.
5. Pearson, Karl. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Phil. Trans.*, 1903, 200-*A*, 1-66.
6. Sichel, H. S. First peace-time validation of army selection tests with a discussion of some statistical problems encountered in this project. *Bulletin of the National Institute for Personnel Research of the South African Council for Scientific and Industrial Research*, 1950, 2, 4-35.
7. Thorndike, R. L. Personnel selection: test and measurement techniques. New York: John Wiley & Sons, 1949.

Manuscript received 10/16/50

PSYCHOMETRIKA—VOL. 16, NO. 3
SEPTEMBER, 1951

COEFFICIENT ALPHA AND THE INTERNAL STRUCTURE OF TESTS*

LEE J. CRONBACH
UNIVERSITY OF ILLINOIS

A general formula (α) of which a special case is the Kuder-Richardson coefficient of equivalence is shown to be the mean of all split-half coefficients resulting from different splittings of a test. α is therefore an estimate of the correlation between two random samples of items from a universe of items like those in the test. α is found to be an appropriate index of equivalence and, except for very short tests, of the first-factor concentration in the test. Tests divisible into distinct subtests should be so divided before using the formula. The index $\bar{r}_{ij}$, derived from α, is shown to be an index of inter-item homogeneity. Comparison is made to the Guttman and Loevinger approaches. Parallel split coefficients are shown to be unnecessary for tests of common types. In designing tests, maximum interpretability of scores is obtained by increasing the first-factor concentration in any separately-scored subtest and avoiding substantial group-factor clusters within a subtest. Scalability is not a requisite.

I. *Historical Resumé*

Any research based on measurement must be concerned with the accuracy or dependability or, as we usually call it, reliability of measurement. A reliability coefficient demonstrates whether the test designer was correct in expecting a certain collection of items to yield interpretable statements about individual differences (25).

Even those investigators who regard reliability as a pale shadow of the more vital matter of validity cannot avoid considering the reliability of their measures. No validity coefficient and no factor analysis can be interpreted without some appropriate estimate of the magnitude of the error of measurement. The preferred way to find out how accurate one's measures are is to make two independent measurements and compare them. In practice, psychologists and educators have often not had the opportunity to recapture their subjects for a second test. Clinical tests, or those used for vocational guidance, are generally worked into a crowded schedule, and there is always a de-

*The assistance of Dora Damrin and Willard Warrington is gratefully acknowledged. Miss Damrin took major responsibility for the empirical studies reported. This research was supported by the Bureau of Research and Service, College of Education.

sire to give additional tests if any extra time becomes available. Purely scientific investigations fare little better. It is hard enough to schedule twenty tests for a factorial study, let alone scheduling another twenty just to determine reliability.

This difficulty was first circumvented by the invention of the split-half approach, whereby the test is rescored, half the items at a time, to get two estimates. The Spearman-Brown formula is then applied to get a coefficient similar to the correlation between two forms. The split-half Spearman-Brown procedure has been a standard method of test analysis for forty years. Alternative formulas have been developed, some of which have advantages over the original. In the course of our development, we shall review those formulas and show relations between them.

The conventional split-half approach has been repeatedly criticized. One line of criticism has been that split-half coefficients do not give the same information as the correlation between two forms given at different times. This difficulty is purely semantic (9, 14); the two coefficients are measures of different qualities and should not be identified by the same unqualified appellation "reliability." A retest after an interval, using the identical test, indicates how stable scores are and therefore can be called a coefficient of *stability*. The correlation between two forms given virtually at the same time, is a coefficient of *equivalence*, showing how nearly two measures of the same general trait agree. Then the coefficient using comparable forms with an interval between testings is a coefficient of equivalence and stability. This paper will concentrate on coefficients of equivalence.

The split-half approach was criticized, first by Brownell (3), later by Kuder and Richardson (26), because of its lack of uniqueness. Instead of giving a single coefficient for the test, the procedure gives different coefficients depending on which items are grouped when the test is split in two parts. If one split may give a higher coefficient than another, one can have little faith in whatever result is obtained from a single split. This criticism is with equal justice applicable to any equivalent-forms coefficient. Such a coefficient is a property of a pair of tests, not a single test. Where four forms of a test have been prepared and intercorrelated, six values are obtained, and no one of these is *the* unique coefficient for Form A; rather, each is the coefficient showing the equivalence of one form to another specific form.

Kuder and Richardson derive a series of coefficients using data from a single trial, each of them being an approximation to the inter-

form coefficient of equivalence. Of the several formulas, one has been justifiably preferred by test workers. In this paper we shall be especially concerned with this, their formula (20):

$$r_{tt(KR20)} = \frac{n}{n-1}\left(1 - \frac{\sum_i p_i q_i}{\sigma_t^2}\right); \quad (i = 1, 2, \cdots n). \tag{1}$$

Here, i represents an item, p_i the proportion receiving a score of 1, and q_i the proportion receiving a score of zero on the item.

We can write the more general formula

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right). \tag{2}$$

Here V_t is the variance of test scores, and V_i is the variance of item scores after weighting. This formula reduces to (1) when all items are scored 1 or zero. The variants reported by Dressel (10) for certain weighted scorings, such as Rights-minus-Wrongs, are also special cases of (2), but for most data computation directly from (2) is simpler than by Dressel's method. Hoyt's derivation (20) arrives at a formula identical to (2), although he draws attention to its application only to the case where items are scored 1 or 0. Following the pattern of any of the other published derivations of (1) (19, 22), making the same assumptions but imposing no limit on the scoring pattern, will permit one to derive (2).

Since each writer offering a derivation used his own set of assumptions, and in some cases criticized those used by his predecessors, the precise meaning of the formula became obscured. The original derivation unquestionably made much more stringent assumptions than necessary, which made it seem as if the formula could properly be applied only to rare tests which happened to fit these conditions. It has generally been stated that α gives a lower bound to "the true reliability"—whatever that means to that particular writer. In this paper, we take formula (2) as given, and make no assumptions regarding it. Instead, we proceed in the opposite direction, examining the properties of α and thereby arriving at an interpretation.

We introduce the symbol α partly as a convenience. "Kuder-Richardson Formula 20" is an awkward handle for a tool that we expect to become increasingly prominent in the test literature. A second reason for the symbol is that α is one of a set of six analogous coefficients (to be designated β, γ, δ, etc.) which deal with such other

concepts as like-mindedness of persons, stability of scores, etc. Since we are concentrating in this paper on equivalence, the first of the six properties, description of the five analogous coefficients is reserved for later publication.

Critical comments on the Kuder-Richardson formula have been primarily directed to the fact that when inequalities are used in deriving a lower bound, there is no way of knowing whether a particular coefficient is a close estimate of the desired measure of equivalence or a gross underestimate. The Kuder-Richardson method is an overall measure of internal consistency, but a test which is not internally homogeneous may nonetheless have a high correlation with a carefully-planned equivalent form. In fact, items within each test may correlate zero, and yet the two tests may correlate perfectly if there is item-to-item correspondence of content.

The essential problem set in this paper is: How shall α be interpreted? α, we find, is the average of all the possible split-half coefficients for a given test. Juxtaposed with further analysis of the variation of split-half coefficients from split to split, and with an examination of the relation of α to item homogeneity, this relation leads to recommendations for estimating coefficients of equivalence and homogeneity.

II. *A Comparison of Split-Half Formulas*

The problem set by those who have worked out formulas for split-half coefficients is to predict the correlation between two equivalent whole tests, when data on two half-tests are at hand. This requires them to define equivalent tests in mathematical terms.

The first definition is that introduced by Brown (2) and by Spearman (33), namely, that we seek to predict correlation with a test whose halves are c and d, possessing data from a test whose halves are a and b, and that

$$V_a = V_b = V_c = V_d; \quad \text{and}$$
$$r_{ab} = r_{ac} = r_{ad} = r_{bc} = r_{bd} = r_{cd}. \tag{3}$$

This assumption or definition is far from general. For many splittings $V_a \neq V_b$, and an equivalent form conforming to this definition is impossible.

A more general specification of equivalence credited to Flanagan [see (25)] is that

$$V_{(a+b)} = V_{(c+d)}; \quad \text{and}$$
$$r_{ab}\sigma_a\sigma_b = r_{ad}\sigma_a\sigma_d = r_{bc}\sigma_b\sigma_c = r_{cd}\sigma_c\sigma_d = \cdots. \tag{4}$$

This assumption leads to various formulas which are collected in the first column of Table 1. All formulas in Column A are mathematically identical and interchangeable.

TABLE 1
Formulas for Split-Half Coefficients

Entering Data*	Formulas Assuming Equal Covariances Between Half-Tests	Formulas Assuming $\sigma_a = \sigma_b$
	1A†	1B‡
$r_{ab}\ \sigma_a\ \sigma_b$	$\dfrac{4\sigma_a\sigma_b r_{ab}}{\sigma_a^2 + \sigma_b^2 + 2\sigma_a\sigma_b r_{ab}}$	$\dfrac{2r_{ab}}{1 + r_{ab}}$
	2A§	
$\sigma_t\ \sigma_a\ \sigma_b$	$2\left(1 - \dfrac{\sigma_a^2 + \sigma_b^2}{\sigma_t^2}\right)$	
	3A‖	
$\sigma_t\ \sigma_a\ r_{at}$	$\dfrac{4(r_{at}\sigma_a\sigma_t - \sigma_a^2)}{\sigma_t^2}$	
	4A¶	4B (≡4A)
$\sigma_t\ \sigma_d$	$1 - \dfrac{\sigma_d^2}{\sigma_t^2}$	$1 - \dfrac{\sigma_d^2}{\sigma_t^2}$
	5A	5B
$\sigma_a\ \sigma_d\ r_{ad}$	$\dfrac{4(\sigma_a^2 - \sigma_a\sigma_d r_{ad})}{4\sigma_a^2 + \sigma_d^2 - 4\sigma_a\sigma_d r_{ad}}$	$\dfrac{2(2\sigma_a^2 - \sigma_d^2)}{4\sigma_a^2 - \sigma_d^2}$

*In this table, a and b are the half-test scores, $t = a+b$, $d = a-b$.
†After Flanagan (25)
‡Spearman-Brown (2, 33)
§Guttman (19)
‖After Mosier (28)
¶Rulon (31)

When a particular split is such that $\sigma_a = \sigma_b$, the Flanagan requirement reduces to the original Spearman-Brown assumption, and in that case we arrive at the formulas in Column B. Formulas 1B and 5B are not identical, since the assumption enters the formulas in different ways. No short formula is provided opposite 2A or 3A, since these exact formulas are themselves quite simple to compute.

Because of the wide usage of Formula 1B, the Spearman-Brown, it is of interest to determine how much difference it makes which assumption is employed. If we divide 1B by any of the formulas in Column A we obtain the ratio

$$k_1 = \frac{2mr + m^2 + 1}{2m(1 + r)} = \frac{1}{(1 + r)}\left(\frac{1 + m^2 + r}{2m}\right), \qquad (5)$$

in which $m = \sigma_b/\sigma_a$, $\sigma_a < \sigma_b$, and r signifies r_{ab}. The ratio when 5B is divided by any of the formulas in the first column is as follows:

$$k_5 = \frac{(2mr - m^2 + 1)(1 + 2mr + m^2)}{2mr(2mr - m^2 + 3)}. \tag{6}$$

When m equals 1, that is, when the two standard deviations are equal, the formula in Column B is identical to that in Column A. As Table 2 shows, there is increasing disagreement between Formula 1B and those in Column A as m departs from unity. The estimate by the Spearman-Brown formula is always slightly larger than the coefficient of equivalence computed by the more tenable definition of comparability.

TABLE 2
Ratio of Spearman-Brown Estimate to More Exact Split-Half Estimate of Coefficient of Equivalence when S.D.'s are Unequal

Ratio of Half-Test S.D.'s (greater/lesser)	Correlation Between Half-Tests					
	.00	.20	.40	.60	.80	1.00
1	1	1	1	1	1	1
1.1	1.005	1.004	1.003	1.003	1.003	1.002
1.2	1.017	1.014	1.012	1.010	1.009	1.008
1.3	1.035	1.029	1.025	1.022	1.020	1.017
1.4	1.057	1.048	1.041	1.036	1.032	1.029
1.5	1.083	1.069	1.060	1.052	1.046	1.042

Formula 5B is not so close an approximation to the results from formulas in Column A. When m is 1.1, for example, the values of k_5 are as follows: for $r = .20, .62$; for $r = .60, .70$; for $r = 1.00, .999$.

It is recommended that the interchangeable formulas 2A and 4A be used in obtaining split-half coefficients. These formulas involve no assumptions contradictory to the data. They are therefore preferable to the Spearman-Brown formula. However, if the ratio of the standard deviations of the half-tests is between .9 and 1.1, the Spearman-Brown formula gives essentially the same result. This finding agrees with Kelley's earlier analysis of much the same question (2, 3).

III. *α as the Mean of Split-Half Coefficients*

To demonstrate the relation between α and the split-half formulas, we shall need the following notation:

Let n be the number of items.

The test t is divided into two half-tests, a and b. i' will designate any item of half-test a, and i'' will designate any item of half-test b. Each half-test contains n' items, where $n' = n/2$.

V_t, V_a, and V_b are the variances of the total test and the respective half-tests.

C_{ij} is the covariance of two items i and j.

C_a is the total covariance for all items in pairs within half-test a, each pair counted once; C_b is the corresponding "within-test" covariance for b.

C_t is the total covariance of all item pairs within the test.

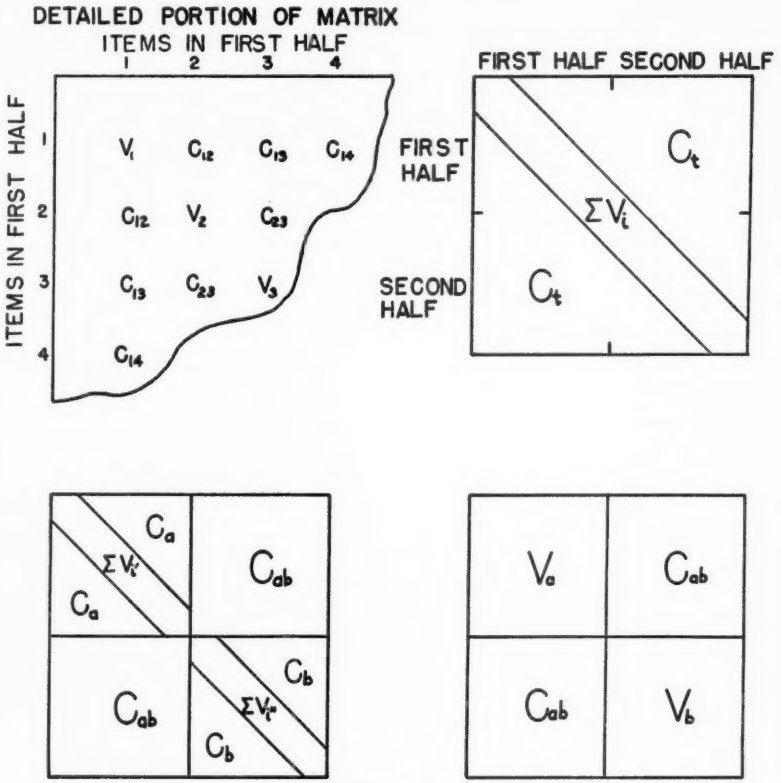


FIGURE 1
Schematic Division of the Matrix of Item Variances and Covariances.

C_{ab} is the total covariance of all item pairs such that one item is within a and the other is within b; it is the "between halves" covariance.

Then

$$C_{ab} = r_{ab}\sigma_a\sigma_b; \tag{7}$$

$$C_t = C_a + C_b + C_{ab}; \tag{8}$$

$$V_t = V_a + V_b + 2C_{ab} = \sum_i V_i + 2C_t; \text{ and} \tag{9}$$

$$V_a = \sum_{i'} V_{i'} + 2C_a \text{ and } V_b = \sum_{i''} V_{i''} + 2C_b. \tag{10}$$

These identities are readily visible in the sketches of Figure 1, which is based on the matrix of item covariances and variances. Each point along the diagonal represents a variance. The sum of all entries in the square is the test variance.

Rewriting split-half formula 2A, we have

$$r_{tt} = 2\left(1 - \frac{V_a + V_b}{V_t}\right) = 2\frac{V_t - V_a - V_b}{V_t}. \tag{11}$$

$$r_{tt} = \frac{4C_{ab}}{V_t}. \tag{12}$$

This indicates that whether a particular split gives a high or low coefficient depends on whether the high interitem covariances are placed in the "between halves" covariance or whether the items having high correlations are placed instead within the same half.

Now we rewrite α:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right) = \frac{n}{n-1}\left(\frac{V_t - \sum_i V_i}{V_t}\right). \tag{13}$$

$$\alpha = \frac{n}{n-1} \cdot \frac{2C_t}{V_t}. \tag{14}$$

$$\bar{C}_{ij} = \frac{C_t}{n(n-1)/2}. \tag{15}$$

Therefore

$$\alpha = \frac{n^2 \bar{C}_{ij}}{V_t}. \tag{16}$$

We proceed now by determining the mean coefficient from all $(2n')!/2(n'!)^2$ possible splits of the test. From (12),

$$\bar{r}_{tt} = \frac{4\,\overline{C}_{ab}}{V_t}. \tag{17}$$

In any split, a particular C_{ij} has a probability of $\frac{n}{2(n-1)}$ of falling into the between-halves covariance C_{ab}. Then over all splits,

$$\Sigma\, C_{ab} = \frac{(2n')!}{2(n'!)^2}\frac{n}{2(n-1)}\sum_i\sum_j C_{ij};\ (i=1,2,\cdots n-1;\ j=i+1,\cdots;n). \tag{18}$$

But

$$\sum_i\sum_j C_{ij} = \frac{n(n-1)}{2}\bar{C}_{ij}. \tag{19}$$

$$\Sigma\, C_{ab} = \frac{(2n')!}{2(n'!)^2}\frac{n^2}{4}\bar{C}_{ij}, \tag{20}$$

and

$$\bar{C}_{ab} = \frac{n^2}{4}\bar{C}_{ij}. \tag{21}$$

From (17),

$$\bar{r}_{tt} = \frac{4n^2}{4V_t}\bar{C}_{ij} = \frac{n^2\,\overline{C}_{ij}}{V_t}. \tag{22}$$

Therefore

$$\bar{r}_{tt} = \alpha. \tag{23}$$

From (14), we can also write α in the form

$$\alpha = \frac{n}{n-1}\frac{\sum_i\sum_j C_{ij}}{V_t};\ (i,j=1,2,\cdots n;\ i\neq j). \tag{24}$$

This important relation states a clear meaning for α as $n/(n-1)$ times the *ratio of interitem covariance to total variance.* The multiplier $n/(n-1)$ allows for the proportion of variance in any item which is due to the same elements as the covariance.

α as a special case of the split-half coefficient. Not only is α a function of all the split-half coefficients for a test; it can also be shown to be a special case of the split-half coefficient.

If we assume that the test is divided into equivalent halves such that $\overline{C}_{i'i''}$ (i.e., C_{ab}/n'^2) equals $\overline{C}_{ij}$, the assumptions for formula 2A still hold. We may designate the split-half coefficient for this splitting as r_{tt_0}.

$$r_{tt} = \frac{4\,C_{ab}}{V_t}. \tag{12}$$

Then

$$r_{tt_0} = \frac{4n'^2\,\overline{C}_{i'i''}}{V_t} = \frac{4n'^2\,\overline{C}_{ij}}{V_t} = \frac{n^2\,\overline{C}_{ij}}{V_t}. \tag{25}$$

From (16),

$$r_{tt_0} = \alpha. \tag{26}$$

This amounts to a proof that α is an exact determination of the parallel-form correlation when we can assume that the mean covariance between parallel items equals the mean covariance between unpaired items. This is the least restrictive assumption usable in "proving" the Kuder-Richardson formula.

α as the equivalence of random samples of items. The foregoing demonstrations show that α measures essentially the same thing as the split-half coefficient. If all the splits for a test were made, the mean of the coefficients obtained would be α. When we make only one split, and make that split at random, we obtain a value somewhere in the distribution of which α is the mean. If split-half coefficients are distributed more or less symmetrically, an obtained split-half coefficient will be higher than α about as often as it is lower than α. This average that is α is based on the very best splits and also on some very poor splits where the items going into the two halves are quite unlike each other.

Suppose we have a universe of items for which the mean covariance is the same as the mean covariance within the given test. Then suppose two tests are made by twice sampling n items at random from this universe without replacement, and administered at the same sitting. Their correlation would be a coefficient of equivalence. The mean of such coefficients would be the same as the computed α. α is therefore an estimate of the correlation expected between two tests drawn at random from a pool of items like the items in this test. Items

are not selected at random for psychological tests where any differentiation among the items' contents or difficulties permits a planned selection. Two planned samplings may be expected to have higher correlations than two random samplings, as Kelley pointed out (25). We shall show that this difference is usually small.

IV. *An Examination of Previous Interpretations and Criticisms of α*

1. *Is α a conservative estimate of reliability?* The findings just presented call into question the frequently repeated statement that α is a conservative estimate or an underestimate or a lower bound to "the reliability coefficient." The source of this conception is the original derivation, where Kuder and Richardson set up a definition of two equivalent tests, expressed their correlation algebraically, and proceeded to show by inequalities that α was lower than this correlation. Kuder and Richardson assumed that corresponding items in test and parallel test have the same common content and the same specific content, i.e., that they are as alike as two trials of the same item would be. In other words, they took the zero-interval retest correlation as their standard. Guttman also began his derivation by defining equivalent tests as identical. Coombs (6) offers the somewhat more satisfactory name "coefficient of precision" for this index which reports the absolute minimum error to be found if the same instrument is applied twice independently to the same subject. A coefficient of stability can be obtained by making the two observations with any desired interval between. A rigorous definition of the coefficient of precision, then, is that it is the limit of the coefficient of stability, as the time between testings becomes infinitesimal.

Obviously, any coefficient of equivalence is less than the coefficient of precision, for one is based on a comparison of different items, the other on two trials of the same items. To put it another way: α or any other coefficient of equivalence treats the specific content of an item as error, but the coefficient of precision treats it as part of the thing being measured. It is very doubtful if testers have any practical need for a coefficient of precision. There is no practical testing problem where the items in the test and only these items constitute the trait under examination. We may be unable to compose more items because of our limited skill as testmakers but any group of items in a test of intelligence or knowledge or emotionality is regarded as a sample of items. If there weren't "plenty more where these came from," performance on the test would not represent performance on any more significant variable.

We therefore turn to the question, does α underestimate appropriate coefficients of equivalence? Following Kelley's argument, the way to make equivalent tests is to make them as similar as possible, similar in distribution of item difficulty and in item content. A pair of tests so designed that corresponding items measure the same factors, even if each one also contains some specific variance, will have a higher correlation than a pair of tests drawn at random from the pool of items. A planned split, where items in opposite halves are as similar as the test permits, may logically be expected to have a higher between-halves covariance than within-halves covariance, and in that case, the obtained coefficient would be larger than α. α is the same type of coefficient as the split-half coefficient, and while it may be lower, it may also be higher than the value obtained by actually splitting a particular test at random. Both the random or odd-even split-half coefficient and α will theoretically be lower than the coefficient from parallel forms or parallel splits.

2. *Is α less than the coefficient of stability?* Some writers expect α to be lower than the coefficient of stability. Thus Guttman says (34, p. 311):

> For the case of scale scores, then, . . . we have the assurance that if the items are approximately scalable [in which case α will be high], then they necessarily have very substantial test-retest reliability.

Guilford says (16, p. 485):

> There can be very low internal consistency and yet substantial or high retest reliability. It is probably not true, however, that there can be high internal consistency and at the same time low retest reliability, except after very long time intervals. If the two indices of reliability disagree for a test, we can place some confidence in the inference that the test is heterogeneous.

The comment by Guttman is based on sound thinking, provided we reinterpret test-retest coefficient on the basis of the context of the comment to refer to the instantaneous retest (i.e., coefficient of precision) rather than the retest after elapsed time. Guilford's statement is acceptable only if viewed as a summary of his experience. There is no mathematical necessity for his remarks to be true. In the coefficient of stability, variance in total score between trials (within persons) is regarded as a source of error, and variance in specific factors (between items within persons) within trials is regarded as true variance. In the coefficient of equivalence, such as α, this is just reversed: variance in specific factors is treated as error. Variation between trials is non-existent and does not reduce true variance (9). Whether the coefficient of stability is higher or lower than the co-

efficient of equivalence depends on the relative magnitude of these variances, both of which are likely to be small for long tests of stable variables. Tests are also used for unstable variables such as mood, morale, social interaction, and daily work output, and studies of this sort are becoming increasingly prominent. Suppose one builds a homogeneous scale to obtain students' evaluations of each day's classwork, the students marking the checklist at the end of each class hour. Homogeneous items could be found for this. Yet the scale would have marked instability from day to day, if class activities varied or the topics discussed had different interest value for different students.

The only proper conclusion is that α may be either higher or lower than the coefficient of stability over an interval of time.

3. *Are coefficients from parallel splits appreciably higher than random-split coefficients or α?* The logical presumption is strong that planned splits as proposed by Kelley (25) and Cronbach (7) would yield coefficients nearer to the equivalent-tests coefficient than random splits do. There is still the empirical question whether this advantage is large enough to be considered seriously. This raises two questions: Is there appreciable variation in coefficients from split to split? If so, does the judgment made in splitting the test into *a priori* equivalent halves raise the coefficient? Brownell (3), Cronbach (8), and Clark (5) have compared coefficients obtained by splitting a test in many ways. There is doubt that the variation among coefficients is ordinarily a serious matter; Clark in particular found that variation from split to split was small compared to variation arising from sampling of subjects.

Empirical evidence. To obtain further data on this question, two analyses were made. One employs responses of 250 ninth-grade boys who took Mechanical Reasoning Test Form A of the Differential Abilities Tests. The second study uses a ten-item morale scale, adapted from the Rundquist-Sletto General Morale Scale by Donald M. Sharpe and administered by him to teachers and school administrators.*

The Mechanical Reasoning Test seems to contain items requiring specific knowledges regarding pulleys, gears, etc. Other items seem to be answerable on the basis of general experience or reasoning. The items seemed to represent sufficiently heterogeneous content that grouping into parallel splits would be possible. We found, however, that items grouped on *a priori* grounds had no higher correlations than items believed to be unlike in content. This finding is con-

*Thanks are expressed to Dr. A. G. Wesman and the Psychological Corporation, and to Dr. Sharpe, for making available the data for the two studies, respectively.

firmed by Air Force psychologists who made a similar attempt to categorize items from a mechanical reasoning test and found that they could not. These items, they note, "are typically complex factorially" (15, p. 309).

Eight items which some students omitted were dropped. An item analysis was made for 50 papers. Using this information, ten parallel splits were made such that items in opposite halves had comparable difficulty. These we call Type I splits. Then eight more splits were made, placing items in opposite halves on the basis of both difficulty and apparent content (Type II splits). Fifteen random splits were made. For all splits, Formula 2A was applied, using the 200 remaining cases. Results appear in Table 3.

TABLE 3

Summary of Data from Repeated Splittings of Mechanical Reasoning Test (60 items; $\alpha = .811$)

	All Splits			Splits Where $1.05 > \sigma_b/\sigma_a > .95$		
Type of Split	No. of Coefficients	Range	Mean	No. of Coefficients	Range	Mean
Random	15	.779–.860	.810	8	.795–.860	.817
Parallel Type I	10	.798–.846	.820	6	.798–.846	.822
Parallel Type II	8	.801–.833	.817	4	.809–.826	.818

There are only 126 possible splits for the morale test, and it is possible to compute all half-test standard deviations directly from the item variances and covariances. Of the 126 splits, six were designated in advance as Type II parallel splits, on the basis of content and an item analysis of a supplementary sample of papers. Results based on 200 cases appear in Table 4.

TABLE 4

Summary of Data from Repeated Splittings of Morale Scale (10 items; $\alpha = .715$)

	All Splits			Splits Where $1.1 > \sigma_b/\sigma_a > .9$		
Type of Split	No. of Coefficients	Range	Mean	No. of Coefficients	Range	Mean
All Splits	126	.609–.797	.715	82	.609–.797	.717
Parallel (Type II)	6	.681–.780	.737	5	.7[illegible]780	.748

The highest and lowest coefficients for the mechanical test differ by only .08, a difference which would be important only when a very precise estimate of reliability is needed. The range for the morale scale is greater (.20), but the probability of obtaining one of the extreme values in sampling is slight. Our findings agree with Clark, that the variation from split to split is less than the variation expected from sample to sample for the same split. The standard error of a Spearman-Brown coefficient based on 200 cases using the same split is .03 when $r_{tt} = .8$, .04 when $r_{tt} = .7$. The former value compares with a standard deviation of .02 for all random-split coefficients of the mechanical test. The standard error of .04 compares with a standard deviation of .035 for the 126 coefficients of the morale test.

This bears on Kelley's comment on proposals to obtain a unique estimate: "A determinate answer would result if the mean for all possible spilts were gotten, but, even neglecting the labor involved, this would seem to contravene the judgment of comparability." (25, p. 79). As our tables show, the splittings where half-test standard deviations are unequal, which "contravene the judgment of comparability," have coefficients about like those which have equal standard deviations.

Combining our findings with those of Clark and Cronbach we have studies of seven tests which seem to show that the variation from split to split is too small to be of practical importance. Brownell finds appreciable variation, however, for the four tests he studied. The apparent contradiction is explained by the fact that the former results applied to tests having fairly large coefficients of equivalence (.70 or over). Brownell worked with tests whose coefficients were much lower, and the larger range of r's does not represent any greater variation in z values at this lower level.

In Tables 3 and 4, the values obtained from deliberately equated half-tests differ slightly, but only slightly, from those for random splits. Where α is .715 for the morale scale, the mean of parallel splits is .748—a difference of no practical importance. One parallel split reaches .780, but this split could not have been defended *a priori* as more logical than the other planned splits. In Table 3, we find that neither Type I nor Type II splits averaged more than .01 higher than α. Here, then, is evidence that the sort of judgment a tester might make on typical items, knowing their content and difficulty, does not, contrary to the earlier opinion of Kelley and Cronbach, permit him to make more comparable half-tests than would be obtained by random splitting. The data from Cronbach's earlier study agree with this. This conclusion seems to apply to tests of any length (the morale scale has

only ten items). Where items fall into obviously diverse subgroups in either content or difficulty, as, say, in the California Test of Mental Maturity, the tester's judgment could provide a better-than-random split. It is dubious whether he could improve on a random division *within subtests.*

It should be noted that in this empirical study no attempt was made to divide items on the basis of r_{it}, as Gulliksen (18, p. 207-210) has recently suggested. Provided this is done on a large sample of cases other than those used to estimate r_{tt}, Gulliksen's plan might indeed give parallel-split coefficients which are consistently at least a few points higher than α.

The failure of the data to support our expectation led to a further study of the problem. We discovered that even tests which seem to be heterogeneous are often highly saturated with the first factor among the items. This forces us not only to extend the interpretation of α, but also to reexamine certain theories of test design.

Factorial composition of the test variance. To make fully clear the relations involved, our analytic procedure will be spelled out in detail. We postulate that the variance of any item can be divided among $k + 1$ orthogonal factors (k common with other items and one unique). Of these, we shall refer to the first, f_1, as the general factor, even though it is possible that some items would have a zero loading on this factor.* Then if f_{zi} is the loading of common factor z on item i,

$$1.00 = N^2(f^2_{1i} + f^2_{2i} + f^2_{3i} + \cdots + f^2_{U_i}). \tag{27}$$

$$C_{ij} = N^2 \sigma_i \sigma_j (f_{1i} f_{1j} + f_{2i} f_{2j} + \cdots + f_{ki} f_{kj}). \tag{28}$$

$$C_t = \sum\sum C_{ij} = N^2 \sum_i \sum_j \sigma_i \sigma_j f_{1i} f_{1j} + \cdots + N^2 \sum_i \sum_j \sigma_i \sigma_j f_{ki} f_{kj};$$

$$(i = 1, 2, \cdots n-1; j = i+1, \cdots, n). \tag{29}$$

$$V_t = N^2 \sum_i \sigma^2_i (f^2_{1i} + \cdots f^2_{ki} + f^2_{U_i}) + 2N^2 \sum_i \sum_j \sigma_i \sigma_j f_{1i} f_{1j}$$

$$+ \cdots + 2N^2 \sum_i \sum_j \sigma_i \sigma_j f_{ki} f_{kj}. \tag{30}$$

If n_1 items contain non-zero loadings on factor 1, and n_2 items contain factor 2, etc., then V_t consists of

*This factor may be a so-called primary or reference factor like Verbal, but it is more likely to be a composite of several such elements which contribute to every item.

n_1^2 terms of the form $N^2\sigma_i\sigma_j f_{1i}f_{1j}$, plus
n_2^2 terms of the form $N^2\sigma_i\sigma_j f_{2i}f_{2j}$, plus (31)
n_3^2 terms of the form $N^2\sigma_i\sigma_j f_{3i}f_{3j}$, plus and so on to
n_k^2 terms of the form $N^2\sigma_i\sigma_j f_{ki}f_{kj}$, plus
n terms of the form $N^2\sigma_i^2 f_{v_i}^2$.

We rarely know the values of the factor loadings for an actual test, but we can substitute values representing different kinds of test structure in (30) and observe the proportionate influence of each factor in the total test.

First we shall examine a test made up of a general factor and five group factors, in effect a test which might be arranged into five correlated subtests. $k = 6$. Let $n_1 = n$, so f_1 is truly general, and let $n_2 = n_3 = n_4 = n_5 = n_6 = 1/5\ n$. To keep the illustration simple, we shall assume that all items have equal variances and that any factor has the same loading (f_z) in all items where it appears. Then

$$\frac{1}{N^2\sigma_i^2} V_t = n^2 f_1^2 + \frac{n^2}{25} f_2^2 + \frac{n^2}{25} f_3^2 + \cdots + \frac{n^2}{25} f_6^2 + \sum_i f_{v_i}^2. \quad (32)$$

It follows that in this particular example, there are n^2 general factor terms, $n^2/5$ group factor terms, and only n unique factor terms. There are, in all, $6n^2/5 + n$ terms in the variance. Let f^2_{zt} be the proportion of test variance due to each factor. Then if we assume that all the terms making up the variance are of the same approximate magnitude,

$$f^2_{1t} = \frac{5n^2}{6n^2 + 5n} = \frac{5n}{6n + 5}. \quad (33)$$

$$\lim_{n\to\infty} f^2_{1t} = \frac{5}{6} = .83. \quad (34)$$

$$f^2_{2t} = \cdots = f^2_{6t} = \frac{n^2/5}{6n^2 + 5n}. \quad (35)$$

$$\lim_{n\to\infty} f^2_{2t} = .03. \quad (36)$$

$$\sum_i f^2_{v_i t} = \frac{5}{6n + 5}. \quad (37)$$

$$\lim_{n\to\infty} \sum_i f^2_{v_i t} = 0. \quad (38)$$

Note that among the terms making up the variance of any test, the number of terms representing the general factor is n times the number representing item specific and error factors.

We have seen that the general factor cumulates a very large influence in the test. This is made even clearer by Figure 2, where we

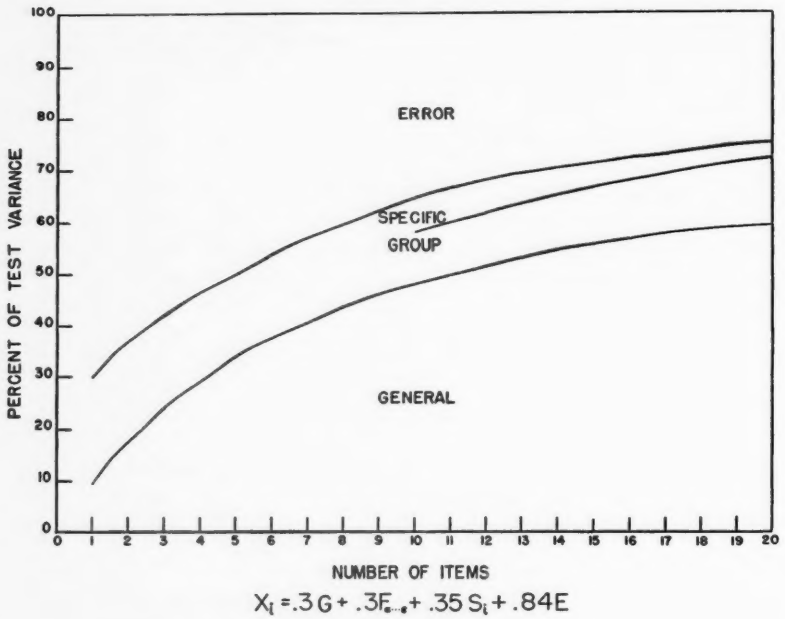


FIGURE 2

Change in Proportion of Test Variance due to General, Group, and Unique Factors among the Items as n Increases.

plot the trend in variance for a particular case of the above test structure. Here we set $k = 6$, $n_1 = n$, $n_2 = n_3 = n_4 = n_5 = n_6 = n/5$. Then we assume that each item has the composition: 9% general factor, 9% from some one group factor, 82% unique. Further, the unique variance is divided by 70/12 between error and specific stable variance. It is seen that even with unreliable items such as these, which intercorrelate only .09 or .18, the general factor quickly becomes the predominant portion of the variance. In the limit, as n becomes indefinitely large, the general factor is 5/6 of the variance, and each group factor is 1/30 of the total variance.

This relation has such important consequences that we work out two more illustrative substitutions in Table 5. We first consider the test which is very heterogeneous in one sense, in that each group of five items introduces a different group factor. No factor save factor 1 is found in more than 5 items. Here great weight in each item is given to the group factor, yet even so, the general factor quickly cumulates in the covariance terms and outweighs the group factors.

The other illustration involves a case where the general factor is much less important in the items than two group factors, each present in half the items. In this type of test, the general factor takes on some weight through cumulation, but the group factors do not fade into insignificance as before. We can generalize that when the proportion of items containing each common factor remains constant as a test is lengthened (factor loadings being constant also), the ratio of the variances contributed by any two common factors remains constant. That is, in such a test pattern each item accounts for a nearly constant fraction of the non-unique variance.

While our description has discussed number of terms, and has simplified by holding constant both item variances and factor loadings, the same general trends hold if these conditions are not imposed. The mathematical notation required is intricate, and we have not attempted a formal derivation of these general principles:

If the magnitude of item intercorrelations is the same, on the average, in successive groups of items as a test is lengthened,

(a) Specific factors and unreliability of responses on single items account for a rapidly decreasing proportion of the variance if the added items represent the same factors as the original items. Roughly, the contribution is inversely proportional to test length.

(b) The ratio in which the remaining variance is divided among the general factor and group factors

  (i) is constant if these factors are represented in the added items to the same extent as in the original items;*

  (ii) increases, if the group factors present in the original items have less weight in the added items.

As a test is lengthened, the general factor accounts for a larger and larger proportion of the total variance. In the case where only a few group factors are present no matter how many items are added,

*This is the case discussed in the recent paper of Guilford and Michael (17). Our conclusion is identical to theirs.

TABLE 5
Factor Composition of Tests Having Certain Item Characteristics as a Function of Test Length

Pattern	Factors	Per Cent of Variance in Any Item	Number of Items Containing Factor	Per Cent of Total Test Variance (assuming equal item variances)				
				$n=1$	$n=5$	$n=25$	$n=100$	$n\to\infty$
One general factor, new group factors in each set of 5 items: $\frac{n}{5}+1 \le k \le \frac{n+4}{5}+1$	f_1	9	n	9	13	44	76	100
	f_2	50	1 to 5	50	74	10	1	0
	.	.	0 to 5			.	.	.
	.	.	"			.	.	.
	.	.	"			.	.	.
	f_k	50 (if present)	"			10	1	0
	$\Sigma f_2 \cdots f_k$			50	74	48	21	0
	f_{U_i}	41	1					
	Σf_{U_i}		n	41	12	8	3	0
				$n=1$	$n=6$	$n=26$	$n=100$	$n\to\infty$
One general factor, two group factors each in half the items	f_1	9	n	9	22	25	26	26
	f_2	50 or 0	$n/2$	50	31	35	36	37
	f_3	0 or 50	$n/2$	0	31	35	36	37
	f_{U_i}	41	1					
	Σf_{U_i}		n	41	17	4	1	0

these also account for an increasing and perhaps substantial portion of the variance. But when each factor other than the first is present in only a few items, the general factor accounts for the lion's share of the variance as the test reaches normal length. We shall return to the implications of this for test design and for homogeneity theory.

Next, however, we apply this to coefficients of equivalence. We may study the composition of half-tests just as we have studied the total test. And we may also examine the composition of C_{ab}, the between-halves covariance. In Table 6, we consider first the test where there is a general factor and two group factors. If the test is divided into halves such that every item is factorially identical to its opposite number, save for the unique factor in each, the covariance C_{ab} nonetheless depends primarily upon the general-factor terms. Note, for example, the twenty-item test. Two-thirds of the covariance terms are the result of item similarity in the general factor. Suppose that these general factor terms are about equal in size. Then, should the test be split differently, the covariance would be reduced to the extent that more than half the items loaded with (say) factor 2 fall in the same half, but even the most drastic possible departure from the parallel split would reduce the covariance by only one-third of its terms. In the event that the group-factor loadings in the items are larger than the general-factor loadings, the size of the covariance is reduced by more than one-third. It is in this case that the parallel split has special advantage: where a few group factors are present and have loadings in the items larger than the general factor does.

The nature of the split has even less importance for the pattern where each factor is found in but a few items. Suppose, for example, that we are dealing with the 60-item test containing 15 factors in four items each. Then suppose that it is so very "badly" split that items containing 5 of the factors were assigned only to one of the half-tests, and items containing the second 5 factors were assigned to the other half-test. This would knock out 40 terms from the between-halves covariance, but such a shift would reduce the covariance only by 40/960 of its terms. Only in the exceptional conditions where general factor loadings are miniscule or where they vary substantially would different splits of such a test produce marked differences in the covariance.

It follows from this analysis that marked variation in the coefficients obtained when a test is split in several ways can result only when

TABLE 6
Composition of the Between-Halves Covariance for Tests of Certain Patterns

Pattern	Common Factors	No. of Items Having Non-Zero Loadings in Factor	No. of Terms Representing Each Factor in Between-Halves Covariance (ΣC_{ij}) When an Ideal Split is Made, for Varying Numbers of Items			
			$n=2$	$n=8$	$n=20$	$n=60$
One general factor,						
two group factors	1	n	1	16	100	900
each in half the	2	$n/2$		4	25	225
items	3	$n/2$		4	25	225
Total No. of Terms in C_{ab}			1	24	150	1350
Total No. of Terms in V_t			8	104	620	5460
			$n=2$	$n=8$	$n=20$	$n=60$
One general factor,	1	n	1	16	100	900
new group factors in	2	4	1	4	4	4
each set of 5 items:	3	4		4	4	4
$\frac{n}{5}+1 \leq k \leq \frac{n+4}{5}+1$	$4 \cdots k$	4 each			4 each	4 each
Total No. of Terms in C_{ab}			2	24	120	960
Total No. of Terms in V_t			10	104	500	3900

(a) a few group factors have substantial loadings in a large fraction of the items or

(b) when first-factor loadings in the items tend to be very small or where they vary considerably. Even these conditions are likely to produce substantial variations only when the variance of a test is contributed to by only a few items.

In the experimental tests studied by Clark, by Cronbach, and in the present study, general-factor loadings were probably greater, on the whole, than group-factor loadings. Moreover, none of the tests seems to have been divisible into large blocks of items each representing one group factor. (Such large "lumps" of group factor content are most often found in tests broken into subtests, viz., the Number Series, Analogies, and other portions of the ACE Psychological examination.)

This establishes on theoretical grounds the fact that for certain common types of test, there is likely to be negligible variation among split-half coefficients. Therefore α, the mean coefficient, represents such tests as well as any parallel split.

This interpretation differs from the Wherry-Gaylord conclusion (38) that "the Kuder-Richardson formula tends to underestimate the true reliability by the ratio $(n - K)/(n - 1)$ when the number of factors, K, is greater than one." They arrive at this by highly restrictive assumptions: that all factors are present in an equal number of items, that no item contains more than one factor, that there is no general factor, and that all items measuring a factor have equal variances and covariances. This type of test would never be intended to yield a psychologically interpretable score. For psychological tests where the intention is that all items include the same factor, our development shows that the quoted statement does not apply.

The problem of differential weighting has been studied repeatedly, the clearest mathematical analyses being those of Richardson (30) and Burt (4). This problem is closely related to our own study of test composition. Making different splits of a test is essentially the same as weighting the component items differently. The conditions under which split-half coefficients differ considerably are identical to those where differential weighting of components alters a total score appreciably: few components, lack of general factor or variation in its loadings, large concentrations of variance in group factors. The more formal mathematical studies of weighting lead to the same conclusions as our study of special cases of test construction.

4. *How is α related to the homogeneity, internal consistency, or*

saturation of a test?* During the last ten years, various writers (12, 19, 27) directed attention to a property they refer to as homogeneity, scalability, internal consistency, or the like. The concept has not been sharply defined, save in the formulas used to evaluate it. The general notion is clear: In a homogeneous test, the items measure the same things.

If a test has substantial internal consistency, it is psychologically interpretable. Two tests, composed of different items of this type, will ordinarily give essentially the same report. If, on the other hand, a test is composed of groups of items, each measuring a different factor, it is uncertain which factor to invoke to explain the meaning of a single score. *For a test to be interpretable, however, it is not essential that all items be factorially similar.* What is required is that a large proportion of the test variance be attributable to the principal factor running through the test (37).

α estimates the proportion of the test variance due to all common factors among the items. That is, it reports how much the test score depends upon general and group, rather than item specific, factors. If we assume that the mean variance in each item attributable to common factors $(\sum_z \sigma_i^2 f_{zi}^2)$ equals the mean interitem covariance $\overline{\sum_z (\sigma_i \sigma_j f_{zi} f_{zj})}$,

$$\frac{1}{n}\sum_z\sum_i \sigma_i^2 f_{zi}^2 = \frac{2}{n(n-1)}\sum_i\sum_j C_{ij} = \frac{2}{n(n-1)} C_t. \qquad (39)$$

$$\sum_z\sum_i \sigma_i^2 f_{zi}^2 = \frac{2}{n-1} C_t, \qquad (40)$$

and the total variance (item variance plus covariance) due to common factors is $2\frac{n}{n-1} C_t$. Therefore, from (14), α is the proportion of test variance due to common factors. Our assumption does not hold true when the interitem correlation matrix has rank higher than one. Normally, therefore, α underestimates the common-factor variance, but not seriously unless the test contains distinct clusters.

The proportion of the test variance due to the first factor among the items is the essential determiner of the interpretability of the

*Several of the comments made in the following sections, particularly regarding Loevinger's concepts, were developed during the 1949 APA meetings in a paper by Humphreys (21) and in a symposium on homogeneity and reliability. The thinking has been aided by subsequent discussions with Dr. Loevinger.

scores. α is an upper bound for this. For those test patterns described in the last section, where the first factor accounts for the preponderance of the common-factor variance, α is a close estimate of first-factor concentration.

α applied to batteries of tests or subtests. Instead of regarding α as an index of *item* consistency, we may apply it to questions of *subtest* consistency. If each subtest is regarded as an "item" composing the test, formula (2) becomes

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum V_{\text{subtests}}}{V_{\text{test}}}\right). \tag{41}$$

Here n is the number of subtests. If this formula is applied to a test or battery composed of separate subtests, it yields useful information about the interpretability of the composite. Under the assumption that the variance due to common factors within each subtest is on the average equal to the mean covariance between subtests, α indicates what proportion of the variance of the composite is due to common factors among the subtests. In many instruments the subtests are positively correlated and intended to measure a general factor. If the matrix of intercorrelations is approximately hierarchical, so that group factors among subtests are small in influence, α is a measure of first-factor concentration in the composite.

Sometimes the variance of the test is not immediately known, but correlations between subtests are known. In this case one can compute covariances ($C_{ab} = \sigma_a \sigma_b r_{ab}$), or the variance of the composite (V_t is the sum of subtest variances and covariances), and apply formula (41). But if subtest variances are not at hand, an inference can be made directly from correlations. If all subtests are assigned weights such that their variances are equal, i.e., they make equal contributions to the total,

$$\alpha = \frac{n}{n-1}\left(\frac{2\sum_i\sum_j r_{ij}}{n + 2\sum_i\sum_j r_{ij}}\right); \; (i = 1, 2, \cdots n-1; j = i+1, \cdots n). \tag{42}$$

Here i and j are subtests, of which there are n. This formula tells what part of the total variance is due to the first factor among the subtests, when the weighted subtest variances are equal.

A few applications will suggest the usefulness of this analysis. The California Test of Mental Maturity, Primary, has two part scores, Language and Non-Language. For a group of 725, according to the

test authors, these scores correlate .668. Then, by (42), α, the common-factor concentration, is .80. Turning to the Primary Mental Abilities Tests, we have a set of moderate positive correlations reported when these were given to a group of eighth-graders (35). The question may be asked: How much would a composite score on these tests reflect common elements rather than a hodgepodge of elements each specific to one subtest? The intercorrelations suggest that there is one general factor among the tests. Computing α on the assumption of equal subtest variances, we get .77. The total score is loaded to this extent with a general intellective factor. Our third illustration relates to four Air Force scores related to carefulness. Each score is the count of number *wrong* on a plotting test. The four scores have rather small intercorrelations (15, p. 687), and each score has such low reliability that its use alone as a measure of carefulness is not advisable. The question therefore arises whether the tests are enough intercorrelated that the general factor would cumulate in a preponderant way in their total. The sum of the six intercorrelations is 1.76. Therefore α is .62. I.e., 62% of the variance in the equally weighted composite is due to the common factor among the tests.

From this approach comes a suggestion for obtaining a superior coefficient of equivalence for the "lumpy" test. It was shown that a test containing distinct clusters of items might have a parallel-split coefficient appreciably higher than α. If so, we should divide the test into subtests, each containing what appears to be a homogeneous group of items. α is computed for each subtest separately by (2). Then $\sigma_i^2\alpha$ gives the covariance of each cluster with the opposite cluster in a parallel form, and the covariance between subtests is an estimate of the covariance of similar pairs "between forms." Hence

$$r_{t_1 t_2} = \frac{\sum_i \sum_j \sigma_i \sigma_j r_{ij}}{V_t}; \; (i = 1, 2, \cdots n; j = 1, 2, \cdots n), \qquad (43)$$

where α_i is entered for r_{ii}, i and j being subtests. To the extent that $\bar{\alpha}_i$ is higher than the mean correlation between subtests, the parallel-forms coefficient will be higher than α_t computed from (2).

The relationships developed are summarized in Figure 3. α falls somewhere between the proportion of variance due to the first factor and the proportion due to all common factors. The blocks representing "other common factors" and "item specifics" are small, for tests not containing clusters of items with distinctive content.

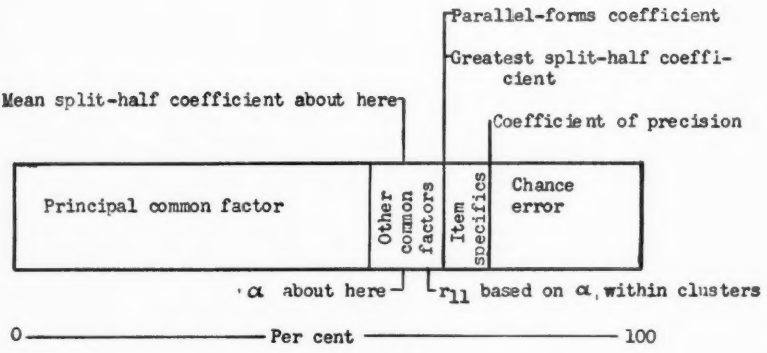


FIGURE 3
Certain Coefficients related to the Composition of the Test Variance.

An index unrelated to test length. Conceptually, it seems as if the "homogeneity" or "internal consistency" of a test should be independent of its length. A gallon of homogenized milk is no more homogeneous than a quart. α increases as the test is lengthened, and so to some extent do the Loevinger-Ferguson homogeneity indices. We propose to obtain an indication of interitem consistency by applying the Spearman-Brown formula to α_t, thereby estimating the mean correlation between items. The formula is entered with the reciprocal of the number of items as the multiple of test length. The formula can be simplified to

$$\bar{r}_{ij(\text{est})} = \frac{\alpha}{n + (1-n)\alpha} \tag{44}$$

or (cf. 24, p. 213 and 30, p. 387),

$$\bar{r}_{ij(\text{est})} = \frac{1}{n-1} \cdot \frac{V_t - \Sigma V_i}{\Sigma V_i}. \tag{45}$$

$\bar{r}_{ij(\text{est})}$ (r bar) is the correlation required, among items having equal variances and equal covariances, to obtain a test of length n having common-factor concentration α. $\bar{r}_{ij(\text{est})}$ or its special case $\bar{\phi}$ for dichotomously-scored items is recommended as an overall index of internal consistency, if one is needed. It is independent of test length. It is not, in my opinion, important for a test to have a high $\bar{r}$ if α is high. Woodbury's "standard length" (39) is an index of internal consistency which can be derived from $\bar{r}_{ij}$ and has the same advantages

and limitations. n_i, the standard length, is the number of items which yields an α of .50. Then

$$n_i = \frac{1 - \bar{r}_{ij}}{\bar{r}_{ij}}. \tag{46}$$

If $\bar{r}$ is high, α is high. But α may be high even when items have small intercorrelations. If $\bar{r}$ is low, the test may be a smooth mixture of items all having low intercorrelations. In this case, each item would have some loading with the general factor and if the test is long α could be high. Such items are illustrated by very difficult psychophysical discriminations such as a series of near-threshold speech signals to be interpreted; with enough of these items we have a highly satisfactory measuring instrument. In fact, save for random error of performance, it may be unidimensional. A low value of $\bar{r}$ may instead indicate a lumpy test composed of discrete and homogeneous subtests. Guttman (34, p. 176n.) describes a questionnaire of this type. The concept of homogeneity has no particular meaning for a "lumpy" test. It is logically meaningless to inquire whether a set of ten measures of physical size plus ten intercorrelated vocabulary items is more homogeneous than twenty slightly correlated biographical questions. A high $\bar{r}$ is sufficient but not necessary evidence that the test lacks important group factors. When $\bar{r}$ is low, only a study of correlations among items or trial clusters of items shows whether the test can be broken into more homogeneous subtests.

Comparison with the index of reproducibility. Guttman's coefficient of reproducibility has appeared to some reviewers (Loevinger, 28; Festinger, 13) as an *ad hoc* index with no mathematical rationale. It may therefore be worthwhile to note that this coefficient can be approximated by a mathematical form which makes clear what it measures. The correlation of any two-choice item with a total score on a test may be expressed as a phi coefficient, and this is common in conventional item analysis. Guttman dichotomizes the test scores at a cutting point selected by inspection of the data. We will get similar results if we dichotomize scores at that point which cuts off the same proportion of cases as pass the item under study. (Our ϕ_{it} will be less in some cases than it would be if determined by Guttman's inspection procedure.) Simple substitution in Guttman's definition (34, p. 117) leads to

$$R \doteq \overline{1 - 2\sigma_i^2(1 - \phi_{it})}, \tag{47}$$

where the approximation is introduced by the difference in ways of

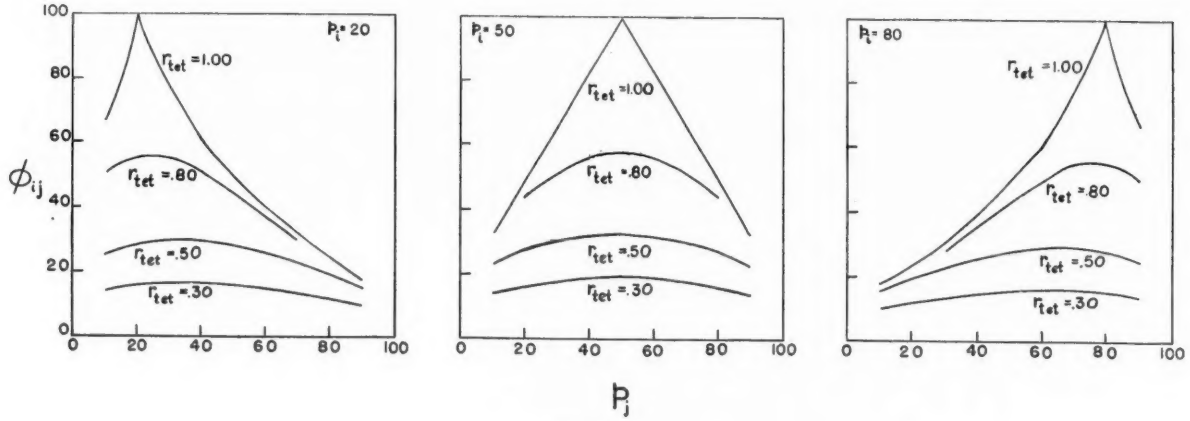


FIGURE 4
Relation of ϕ_{ij} to p_i and p_j for Several Levels of Correlation.

dichotomizing. The actual R obtained by Guttman will be larger than that from (47). For multiple-alternative items, a similar but more complex formula involving the phi coefficient of the alternative with the test is required to approximate Guttman's result. R is independent of test length; if a Guttman scale is divided into equivalent portions, the two halves will have the same R as the original test. In this respect, R is most comparable to our $\bar{r}$. Both ϕ_{it} and $\bar{r}$ are low, so long as items are unreliable or contain substantial specific factors.

5. *Is the usefulness of α limited by properties of the phi coefficient between items having unequal difficulties?* The criticism has been made, most vehemently by Loevinger (27), that α is a poor index because, being based on product-moment correlations, it cannot attain unity unless all items have distributions of the same shape. For the pass/fail item, this requires that all p_i be equal. The inference is drawn that since the coefficient cannot reach unity for such items, α and $\bar{r}$ do not properly represent homogeneity.

There are two ways of examining this criticism. The simpler is empirical. The alleged limitation upon the product-moment coefficient has no practical effect upon the coefficient, for items of the sort customarily employed in psychological tests. To demonstrate this, we consider the change in ϕ with changes in item difficulty. To hold constant the relation between the "underlying traits," we fix the tetrachoric correlation. When the tetrachoric coefficient is .30, $p_i = .50$ and p_j ranges from .10 to .90, ϕ_{ij} ranges only from .14 to .19. Figure 4 shows the relation of ϕ_{ij} to p_i and p_j for three levels of correlation: $r_{\text{tet}} = .30$, $r_{\text{tet}} = .50$, and $r_{\text{tet}} = .80$. The correlation among items in psychometric tests is ordinarily below .30. For example, even for a five-grade range of talent, the $\overline{\phi}_{ij}$ for the California Test of Mental Maturity subtests range only from .13 to .25. That is, for tests having the degree of item intercorrelation found in present practice, ϕ is very nearly constant over a wide range of item difficulties.

TABLE 7

Variation in Certain Indices of Interitem Consistency with Changes in Item Difficulty (Tetrachoric Correlation Held Constant)

p_i	.50	.50	.50	.50	.50	.50	.50	.50	.50
p_j	→.00	.10	.20	.40	.50	.60	.80	.90	→1.00
$r_{ij\text{tet}}$	.30	.30	.30	.30	.30	.30	.30	.30	.30
ϕ_{ij}	→.00	.14	.17	.19	.19	.19	.17	.14	→.00
H_{ij}	→1.00	.42	.34	.23	.19	.23	.34	.42	→1.00

Examining Loevinger's proposed coefficient of homogeneity (29),

$$H_{ij} = \phi_{ij}/\phi_{ij(\max)}\,, \tag{48}$$

we find that *it* is markedly affected by variations in item difficulty. One example is worked out in Table 7. As many investigators including Loevinger have noted, Guttman's R is drastically affected by item difficulty. For any single item, R must be greater than p_i or q_j, whichever is greater. Evidently the indices of homogeneity which might replace $\overline{\phi}$ suffer more from the effects of differences in difficulty than does the phi coefficient.

Further evidence on the alleged limitation of α is obtained by preparing four hypothetical 45-item tests. In each case, all $r_{ij(\text{tet})}$ are fixed at .30. Phi coefficients reflect both heterogeneity in content and heterogeneity in difficulty. To assess the effect of the latter heterogeneity upon $\overline{\phi}$ and α, we compared one test of uniform item difficulty, where all heterogeneity is in content, with another where "heterogeneity due to difficulty" was allowed to enter. As Table 8 indicates, even when extreme ranges of item difficulty are allowed, neither $\overline{\phi}$ nor α is affected in any practically important way. For tests where item difficulties are higher, or correlations are lower, the effect would be even more negligible.

TABLE 8

Comparison of $\overline{\phi}$ and α for Hypothetical 45-Item Tests With and Without "Heterogeneity Due to Item Difficulty"

Test	Distribution of Difficulties	Range of p_i	$\overline{p_i}$	$\overline{\phi}$	Diff.	α	Diff.
A	Normal	.20 to .80	.50	.181	*.011*	.909	*.005*
A′	Peaked	.50	.50	.192		.914	
B	Normal	.10 to .90	.50	.176	*.016*	.906	*.008*
B′	Peaked	.50	.50	.192		.914	
C	Normal	.50 to .90	.70	.170	*.011*	.902	*.007*
C′	Peaked	.70	.70	.181		.909	
D	Rectangular	.10 to .90	.50	.153	*.039*	.892	*.022*
D′	Peaked	.50	.50	.192		.914	

Still another small study leading to the same essential conclusion was made by examining a "perfect scale," where all p_{ij} equal $\phi_{ij(\max)}$. Items were placed at five difficulty levels, the p_i being .50, .58, .71, .80, and .89. Then the correlations (phis) of items range from 1.00 (at same level) to .85 (highest between levels) to .36. In a test of

only five items, α reaches .86. This is the maximum α could have, for this set of 5 items and specified p_i. As the number of items increases, α rises toward 1.00. Thus, for 10 items, two at each level $\alpha_{max} = .951$; for 20 items, .977. It follows that even if items are much more homogeneous in content than present tests and much freer from error, the cumulative properties of covariance terms make the failure of all ϕ's to reach unity of next-to-no importance. α_{max} would be lower if difficulties range over the full scale, but the same principle holds. α is a good measure of common-factor concentration, for tests of reasonable length, in spite of the fact that it falls short of 1.00 if items vary in difficulty.

In the case of the perfect scale, of course, $\bar{\phi}$ does fall well short of unity and for such tests it does not reflect the homogeneity in content. From the five-item case just considered, $\bar{\phi}$ is .54.

The second way to analyze this criticism is to examine the nature of redundancy (using a term from Shannon's information theory, 32). If two items repeat the same information, they are totally redundant. Thus, if one item divides people 50/50, and the second item does also, the two items always placing exactly the same people together, the second item gives no new information about individual differences. (Cf. Tucker, 36). Suppose, though, that the second item is passed by 60 per cent of the subjects. Even if $r_{ij(tet)} = 1.00$, this second item conveys new information because it discriminates among the fifty people who failed the first item. A five-item test where all items have perfect tetrachoric intercorrelations, and the p_i are .40, .45, .50, .55, .60, is perfectly homogeneous (a la Guttman, Loevinger, et al). So is a ten-item test composed of these items plus five others whose p's are .30, .35, .65, .70, .75. The two tests are not equivalent in measuring power, however; the second makes a much greater number of discriminations. Because there is less redundancy, the longer test has a lower $\bar{\phi}$.

From the viewpoint of information theory, we should be equally concerned with heterogeneity in content and heterogeneity in difficulty. We get one bit of information when we place the person as above the mean in (say) pitch discrimination. Now with another item or set of items, we might place him relative to the mean in visual acuity. The two tests together place him in one of four categories. If our second test had been a further measure of pitch, placing the subject above or below the 75th percentile, then the two tests would have placed him in one of four categories. Either set of tests gives the same amount of information. Which information we most want de-

pends on practical considerations.

The phi coefficient reports whether a second item gives new information that the first does not. Then a tetrachoric r must be computed to determine if the new information relates to a new content dimension or to a finer discrimination on the same content dimension. If the phi coefficient between true scores is 1.00, redundancy is complete and there is no new information. Redundancy is desirable when accuracy of a single item is low. To test whether men can hear a 10-cycle difference, the best way is to use a large number of items of just that difficulty. Such items usually also discriminate to some degree at other points on the scale, but cannot give information about ability at the 5-cycle level if a single item is extremely reliable. With very accurate items a pitch test which is not homogeneous will be better for differentiation all along the scale. The "factors" found by Ferguson (11) due to the higher correlation (redundancy) of items with equal difficulty need not be regarded as artifacts (38).* These "difficulty factors" are factors on which the test gives information and on which the tester may well want information. They are not "content factors," but they must be considered in test analysis. For example, if one regards pitch tests in this light, it is seen that a test containing 5-cycle items, 10-cycle items, and 15-cycle items will be slightly influenced by undesired factors, when the criterion requires discrimination only at the 15-cycle level. (Problems of this type occur in validating tests for selecting military personnel using detection apparatus). One would maximize the loading in the test of the group factor among 15-cycle items, to maximize validity. This factor is of course a mathematical factor, and not a property of the auditory machinery. While the mathematics is not clear, it seems very likely that the group factors found among phi coefficients are interchangeable with Guttman's "components of scale analysis" to which he gives serious psychological interpretation.

From this point of view, the phi coefficient which tells when items do and do not duplicate each other is a better index *just because* it does not reach unity for items of unequal difficulty. Phi and r_{tet} are both useful in test analysis. Brogden (1, pp. 199, 201) makes a similar point, although approaching the problem from another tack.

*It is not *necessary*, as Ferguson seems to think, for difficulty factors to emerge if product-moment correlations are used with multi-category variates. On *a priori* grounds, difficulty factors will appear only if the shapes of the distributions of the variates are different. In Ferguson's data it appears likely that the hardest and easiest tests were skewed in opposite directions.

Implications for Test Design

In view of the relations detailed above, we find it unnecessary to create homogeneous scales such as Guttman, Loevinger, and others have urged.

It is true that a test where all items represent the same content factor with no error of measurement is maximally interpretable. Everyone attaining the same score would mark items in the same way. Yet the question we really wish to ask is whether the individual differences in test score are attributable to the first factor within the test. If a large proportion of the score variance relates to this factor, the residue due to specific characteristics of the items little handicaps interpretability. It has been shown that a high first-factor saturation indicated by a high α can be attained by cumulating many items which have low correlations. The standard proposed by Ferguson, Loevinger, and Guttman is unreasonably severe, since it would rule out tests which do have high first-factor concentrations.

These writers seem to wish to infer the person's score on each item from his total score. This appears unimportant, but even if it were important, the interest would attach to predicting his *true* standing on the item, not his fallible obtained score. For the unreliable items used in psychological and educational tests, the aim of Guttman et al. will not be approached in practice. Perhaps sociological data have such greater reliability that prediction of obtained scores is tantamount to predicting true scores.

Increasing interpretability by lengthening a test is not without its disadvantages. Using more and more time to get at the same information employs the principle of redundancy (32). When a message is repeated over and over, it is easier to infer the true message even when there is substantial interference (item unreliability). But the more you repeat messages already transmitted, the less time is allowed for conveying other information. A set of redundant items can carry much less information than a set of independent items. In other words, when we lengthen certain tests or subtests to make their scores more interpretable, we sacrifice the possibility of obtaining separate measures of additional factors in the same time.

From the viewpoint of both interpretability and efficient prediction of criteria, the smallest element on which a score is obtained should be a set of items having a substantial α and not capable of division into discrete item clusters which themselves have high α. Such separately interpretable tests can sometimes be combined into an interpretable composite, as in the case of the PMA tests. Although

it is believed that the test designer should seek interitem consistency, and judge the effectiveness of his efforts by the coefficient α, the pure scale should not be viewed as an ideal. It should be remembered that Tucker (36) and Brogden (1) have demonstrated that increases in internal consistency may lead to decreases in the product-moment validity coefficient when the shape of the test-score distribution differs from that of the criterion distribution.

Summary

1. Formulas for split-half coefficients of equivalence are compared, and those of Rulon and Guttman are advocated for practical use rather than the Spearman-Brown formula.

2. α, the general formula of which Kuder-Richardson formula 20 is a special case, is found to have the following important meanings:

(a) α is the mean of all possible split-half coefficients.

(b) α is the value expected when two *random* samples of items from a pool like those in the given test are correlated.

(c) α is a lower bound for the coefficient of precision (the instantaneous accuracy of this test with these particular items). α is also a lower bound for coefficients of equivalence obtained by simultaneous administration of two tests having matched items. But for reasonably long tests not divisible into a few factorially-distinct subtests, α is nearly equal to "parallel-split" and "parallel-forms" coefficients of equivalence.*

(d) α estimates, and is a lower bound to, the proportion of test variance attributable to common factors among the items. That is, it is an index of common-factor concentration. This index serves purposes claimed for indices of homogeneity. α may be applied by a modified technique to determine the common-factor concentration among a battery of subtests.

*W. G. Madow suggests that the amount of disagreement between two random or two planned samples of items from a larger population of items could be anticipated from sampling theory. The person's score on a test is a sample mean, intended to estimate the population mean or "true score" over all items. The variance of such a mean from one sample to another decreases rapidly as the sample is enlarged by lengthening the test, whether samples are drawn at random or are drawn after stratifying the universe as to difficulty and content. The conditions under which the radom splits correlate about as highly as parallel splits are those in which stratified sampling has comparatively little advantage. Madows comment has implications also for the preparation of comparable forms of tests and for developing objective methods of selecting a sample of items to represent a larger set of items so that the variance of the difference between the score based on the sample and the score based on the universe of items is as small as possible.

(e) α is an upper bound to the concentration in the test of the first factor among the items. For reasonably long tests not divisible into a few factorially-distinct subtests, α is very little greater than the exact proportion of variance due to the first factor.

3. Parallel-splits yield coefficients little larger than random splits, unless tests contain large blocks of items representing group factors. For such tests, α computed for separate blocks and combined by a special formula gives a satisfactory estimate of first-factor concentration.

4. Interpretability of a test score is enhanced if the score has a high first-factor concentration. A high α is therefore to be desired, but a test need not approach a perfect scale to be interpretable. Items with quite low intercorrelations can yield an interpretable scale.

5. A coefficient $\bar{r}_{ij}$ (or $\bar{\phi}_{ij}$) is derived which is the intercorrelation required, among items with equal intercorrelations and variances, to reproduce a test of n items having common-factor concentration α. $\bar{\phi}$, as a measure of item interdependence, draws attention to heterogeneity in both difficulty and content factors. Heterogeneity in test difficulty merits the attention of the test designer, since the validity of the test may be increased by capitalizing on "difficulty factors" present in the criterion.

6. To obtain subtest scores for interpretation or to be weighted in an empirical composite, the ideal set of items is one having a substantial α and not further divisible into a few discrete smaller blocks of items.

REFERENCES

1. Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 1946, **11**, 197-214.
2. Brown, W. Some experimental results in the correlation of mental abilities. *Brit. J. Psychol.*, 1910, **3**, 296-322.
3. Brownell, W. A. On the accuracy with which reliability may be measured by correlating test halves. *J. exper. Educ.*, 1933, **1**, 204-215.
4. Burt, C. The influence of differential weighting. *Brit. J. Psychol., Stat. Sect.*, 1950, **3**, 105-128.
5. Clark, E. L. Methods of splitting vs. samples as sources of instability in test-reliability coefficients. *Harvard educ. Rev.*, 1949, **19**, 178-182.
6. Coombs, C. H. The concepts of reliability and homogeneity. *Educ. psychol. Meas.*, 1950, **10**, 43-56.
7. Cronbach, L. J. On estimates of test reliability. *J. educ. Psychol.*, 1943, **34**, 485-494.

8. Cronbach, L. J. A case study of the split-half reliability coefficient. *J. educ. Psychol.*, 1946, **37**, 473-480.
9. Cronbach, L. J. Test "reliability": its meaning and determination. *Psychometrika*, 1947, **12**, 1-16.
10. Dressel, P. L. Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika*, 1940, **5**, 305-310.
11. Ferguson, G. The factorial interpretation of test difficulty. *Psychometrika*, 1941, **6**, 323-329.
12. Ferguson, G. The reliability of mental tests. London: Univ. of London Press, 1941.
13. Festinger, L. The treatment of qualitative data by "scale analysis." *Psychol. Bull.*, 1947, **44**, 149-161.
14. Goodenough, F. L. A critical note on the use of the term "reliability" in mental measurement. *J. educ. Psychol.*, 1936, **27**, 173-178.
15. Guilford, J. P., ed. Printed classification tests. Report No. 5, Army Air Forces Aviation Psychology Program. Washington: U. S. Govt. Print. Off., 1947.
16. Guilford, J. P. Fundamental statistics in psychology and education. Second ed. New York: McGraw-Hill, 1950.
17. Guilford, J. P., and Michael, W. B. Changes in common-factor loadings as tests are altered homogeneously in length. *Psychometrika*, 1950, **15**, 237-249.
18. Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.
19. Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, **10**, 255-282.
20. Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, **6**, 153-160.
21. Humphreys, L. G. Test homogeneity and its measurement. *Amer. Psychologist*, 1949, **4**, 245.
22. Jackson, R. W., and Ferguson, G. A. Studies on the reliability of tests. Bull. No. 12, Dept. of Educ. Res., University of Toronto, 1941.
23. Kelley, T. L. Note on the reliability of a test: a reply to Dr. Crum's criticism. *J. educ. Psychol.*, 1924, **15**, 193-204.
24. Kelley, T. L. Statistical method. New York: Macmillan, 1924.
25. Kelley, T. L. The reliability coefficient. *Psychometrika*, 1942, **7**, 75-83.
26. Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, **2**, 151-160.
27. Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, **61**, No. 4.
28. Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychol. Bull.*, 1948, **45**, 507-529.
29. Mosier, C. I. A short cut in the estimation of split-halves coefficients. *Educ. psychol. Meas.*, 1941, **1**, 407-408.
30. Richardson, M. Combination of measures, pp. 379-401 in Horst, P. (Ed.) The prediction of personal adjustment. New York: Social Science Res. Council, 1941.
31. Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. *Harvard educ. Rev.*, 1939, **9**, 99-103.
32. Shannon, C. E. The mathematical theory of communication. Urbana: Univ. of Ill. Press, 1949.

33. Spearman, C. Correlation calculated with faulty data. *Brit. J. Psychol.*, 1910, 3, 271-295.
34. Stouffer, S. A., et. al. Measurement and prediction. Princeton: Princeton Univ. Press, 1950.
35. Thurstone, L. L., and Thurstone, T. G. Factorial studies of intelligence, p. 37. Chicago: Univ. of Chicago Press, 1941.
36. Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika*, 1946, 11, 1-13.
37. Vernon, P. E. An application of factorial analysis to the study of test items. *Brit. J. Psychol., Stat. Sec.*, 1950, 3, 1-15.
38. Wherry, R. J., and Gaylord, R. H. The concept of test and item reliability in relation to factor pattern. *Psychometrika*, 1943, 8, 247-264.
39. Woodbury, M. A. On the standard length of a test. Res. Bull. 50-53, Educ. Test. Service, 1950.

Manuscript received 12/28/50
Revised manuscript received 2/28/51

PSYCHOMETRIKA—VOL. 16, NO. 3
SEPTEMBER, 1951

ESTIMATION OF OTHER COEFFICIENTS OF CORRELATION FROM THE PHI COEFFICIENT

J. P. GUILFORD AND NORMAN C. PERRY
UNIVERSITY OF SOUTHERN CALIFORNIA

Formulas are developed for estimating a point-biserial r or a tetrachoric r from an obtained phi coefficient. The estimate of a tetrachoric r, which is called r_ϕ, is shown to be equivalent to that obtained from first-order use of the tetrachoric r series. A tabulation is made of corrections needed to make r_ϕ equivalent numerically to the tetrachoric r. In spite of its greater generality than estimates of tetrachoric r by previous methods, there are limitations, which are pointed out.

In recent years there has been an increasing recognition of the utility of the Boas-Yule phi coefficient as an index of correlation. This is evident in the greater attention it receives in general textbooks on statistical methods (2, 3, 5). Some of this attention is due to the increased importance of categorical data and the more extensive, rigorous use of them in research in the social sciences. Some is consistent with the view that responses to test items should be regarded operationally as categorical data calling for corresponding treatment in item analysis.

Needs for Estimation of Other Indices of Correlation from Phi

We first mention some examples of occasions for estimating other types of correlation coefficients when only a phi coefficient is known.

There are a number of circumstances under which this happens. Data in which both variables come to us dichotomized yield a phi coefficient with little computation. We may decide that both of the variables are actually continuous, the regression is rectilinear, and distributions of continuous measurements of an appropriate type would be normal in the population. A tetrachoric correlation coefficient would be obviously called for, as the equivalent of the Pearson r. But perhaps we have already computed a phi coefficient, or we prefer the latter for computational reasons. Can we estimate the tetrachoric r from the phi coefficient with sufficiently tolerable accuracy to justify its use in this situation?

Another situation arises in connection with the computation of intercorrelations by means of the IBM tabulator. The distributions of some of the variables may be so skewed or otherwise irregular that Pearson r's would be out of the question. Such irregularities are likely to introduce non-linearity or heteroscedasticity or both (3, p, 171). Yet, it may well be assumed that the population distributions on ideal scales would be normal. One practical solution would be to code the scores in such distributions as 1 (above-median) and 0 (below-median). The product-moment correlation of two such variables gives a phi coefficient. Yet it is a value equivalent to a Pearson r that we want, to represent the genuine degree of relationship.

In item-analysis studies, it is quite common procedure, for convenience, to dichotomize the total-score distribution and to correlate with it the pass-fail dichotomy for each item. From the four-fold table we may compute either a tetrachoric r or a phi coefficient. The tetrachoric r is called for if we want to know how well the thing (or things) measured by the item correlates with the thing (or things) measured by the total score. If we want to know how well we can predict total score from the item dichotomy, however, the coefficient we want is a point-biserial r (3, p. 499). In other words the point-biserial r is the realistic one for determining how well the item functions in measurement. The realistic correlation coefficient for correlations between items would be the phi coefficient, if we want to know how well we could predict success on one item from success on another item. If we wanted to make a factor analysis based upon the intercorrelations of items, however, we would want tetrachoric r's.

Having computed phi coefficients, either between items and total scores or between pairs of items, can we make tolerably accurate estimates of either point-biserial r's or tetrachoric r's? If we can, the great economy of computing phi in the item-analysis situation gives it considerable appeal (3, p. 503). From it we could derive an estimate of either the point-biserial r or the tetrachoric r.

Common Correction Practices

The procedures for estimating a tetrachoric r from phi as given in the textbooks have been unsatisfactory. The problem has been inadequately approached from the point of view of correcting a coefficient of correlation for coarse grouping. Peatman (5) recommends the division of ϕ by the constant .798 to estimate a point-biserial and by the constant .637 (which is $.798^2$) to estimate a tetrachoric r. Guilford (3), makes a similar recommendation, with serious reservations.

These constants are called for under the general principle of correction for coarse grouping when the category values are the means of the cases in the categories. It will be shown that these constants actually apply only under one set of conditions. Edwards (2) follows the approximation procedure given by Camp (1), which varies the correction constant according to the size of p, the proportion of cases in the category with the larger marginal frequency.

A General Correction Formula

We will now proceed to develop a formula for the estimation of a Pearson r from a phi coefficient and to show its relation to the series for a tetrachoric r and its limits of accuracy. The basis will be the principle of correcting r for coarse grouping when there are two categories in X and Y, and when the index numbers are the means of values in segments of the distributions. It will be necessary to assume normal, continuous distributions in X and Y.

According to the usual mathematical development, which is given by Peters and Van Voorhis (6, p. 395), the correction factor for either variable is the standard deviation of the grouped values. Let us assume a unit normal distribution for X and for Y. Some needed parameters of either distribution will be:

p' = the proportion of the N cases in the upper category of X,

p = the similar proportion on Y,

q' = the proportion of the N cases in the lower category on X,

q = the similar proportion on Y,

y' = the ordinate at the point of division on distribution X, and

y = the similar ordinate in distribution Y.

It is necessary, next, to express the standard deviations in terms of these parameters. With the distribution normal, the means of the two segments are given by the two ratios, y/p and y/q, for the distribution on Y. The standard deviation of the distribution on Y is given by the equation

$$\sigma_y = \sqrt{\frac{p\left(\frac{y}{p}\right)^2 + q\left(\frac{y}{q}\right)^2}{p+q}}$$

$$= \sqrt{\frac{y^2}{p} + \frac{y^2}{q}}$$

$$= \sqrt{\frac{qy^2 + py^2}{pq}}$$

$$= y\sqrt{\frac{p+q}{pq}}$$

$$\sigma_y = \frac{y}{\sqrt{pq}}. \tag{1}$$

In a similar manner it can be shown that

$$\sigma_x = \frac{y'}{\sqrt{p'q'}}. \tag{2}$$

Applying both corrections to phi,

$$r_\phi = \frac{\phi}{\left(\frac{y'}{\sqrt{p'q'}}\right)\left(\frac{y}{\sqrt{pq}}\right)}, \tag{3}$$

in which r_ϕ is an estimate of a Pearson r from a known ϕ coefficient. It will, therefore, be seen that the correction factors are functions of p and p', the marginal proportions. These functions are at a minimum when $p = .5$. *Under this condition, a correction factor equals .798, and only under the condition that* $p = p' = .5$ *will the joint correction factor be .637.*

For greater convenience, it is better in practice to use the reciprocals of the standard deviations. Equation (3) then becomes

$$r_\phi = \phi \frac{(\sqrt{p'q'})}{y'} \frac{(\sqrt{pq})}{y}. \tag{4}$$

Tabled values of these correction factors are available (3, p. 614).

A point-biserial r would be estimated from phi by the use of

only one of these multiplying factors, since there is only one genuinely continuous distribution. A Pearson r could be estimated from a computed point-biserial r by multiplying the latter by one of these same factors. As a matter of fact, such a factor was previously shown to be the ratio of proportionality between a biserial r and a point-biserial r computed from the same data (3, p. 331), by relating the two formulas for those two coefficients.

TABLE 1

Correction Factors for Estimating r_t from ϕ for Different Combinations of p and p'

	$\sqrt{p'q'}$	p				
p'	y'	.5	.6	.7	.8	.9
.5	1.253	1.570	1.589	1.651	1.791	2.141
.6	1.268		1.608	1.671	1.812	2.167
.7	1.318			1.737	1.883	2.252
.8	1.429				2.042	2.442
.9	1.709					2.921

To give some conception of how much error is introduced in using the constant .637 regardless of the values of p and p' Table 1 is presented. Here the correction factors of formula (4) corresponding to p values of .5, .6, .7, .8, and .9 are given, also the products of those factors for various combinations of p and p'. For the combination of $p = .5$ and $p' = .5$, the factor is 1.570, which is the reciprocal of .637. For the most extreme case in the table, when $p = .9$ and $p' = .9$, the multiplier equals 2.921, which is almost twice that for the minimum product. The procedure suggested by Camp takes into account only one of the p values, whichever is larger. On that basis, his corrections vary much less with p than do those by the method proposed here.

Let us substitute in equation (4) the usual expression for ϕ that is used in its computation. We then have

$$r_\phi = \left(\frac{a\delta - \beta\gamma}{\sqrt{pqp'q'}} \right) \left(\frac{\sqrt{pqp'q'}}{yy'} \right)$$

or

$$r_\phi = \frac{a\delta - \beta\gamma}{yy'}. \qquad (5)$$

We thus arrive at what seems to be a new equation for estimating directly from a fourfold table of proportions the Pearson coefficient of correlation.

If we start with the fourfold table of frequencies instead of proportions, formula (5) becomes

$$r_\phi = \frac{ad - bc}{yy' N^2}. \tag{6}$$

Equation (6) will look familiar to many, for the right-hand term is precisely the left-hand side of the text-book infinite series equation for tetrachoric r, (3, p. 334).*

$$\frac{ad - bc}{yy' N^2} = r_t + r_t^2 \frac{hk}{2} + r_t^3 \frac{(h^2 - 1)(k^2 - 1)}{6} + \cdots$$

If the tetrachoric r is the best estimate of the Pearson r that can be obtained from a 2×2 table, then the amount of error involved in using the formulas given here in estimating the Pearson r is indicated by the contribution of all terms of the infinite series with powers higher than the first. For small and even moderate values of r_t, there is probably little error in so doing. When r_t becomes large, however, the error becomes serious. We next turn our attention to estimating this discrepancy under different conditions.

Refinement of the Estimation

This error, say $\varepsilon = r_t - r_\phi$, can be conveniently obtained by iterative processes if r_ϕ is less than .6, and the proportions in the dichotomies are not too far from .5. To illustrate the iteration let us solve for r_t for the data in Table 2.

TABLE 2
An Illustrative Problem

		Question I: No	Question I: Yes	Total	Proportion
Question II	Yes	167	374	541	.582
	No	203	186	389	.418
	Total	370	560	930	
	Prop.	.398	.602		

*We are indebted to Mr. Russel F. Green for pointing out this identity.

The required infinite series* is given by the equation

$$r_\phi = r_t + r_t^2\left(\frac{hk}{2}\right) + r_t^3\frac{(h^2-1)(k^2-1)}{6} + \cdots \qquad (7)$$

where h and k are the z-scores of the points of division determining the dichotomies. In practice the coefficients for the polynomial in r_t are most conveniently obtained from Pearson's Tables (4).

For the data of Table 2, then, equation (7) becomes

$$\frac{(374)(203)-(167)(186)}{(.3905)(.3858)(930)^2} = r_t + \frac{(.2070)(.2585)}{2}r_t^2 + \cdots$$

$$\frac{44860}{(.1570)(864900)} = .3443 = r_t + .02675r_t^2\frac{(.9322)(.9572)}{6}r_t^3 + \cdots$$

$$r_t = .3443 - .02675r_t^2 - .1489r_t^3 - .0193r_t^4 - .0595r_t^5 - .0155r_t^6$$

where $h = .2070$ and $k = .2585$.

As a first approximation to r_t, take $r_1 = .3443$ and substitute on the right. Take the new result r_2 and again substitute on the right obtaining r_3 etc.

$$r_2 = .3443 - .0032 - .0061 - .0003 - .0003 - \cdots = .3344$$
$$r_3 = .3443 - .0030 - .0056 - .0002 - .0003 - \cdots = .3352$$
$$r_4 = .3443 - .0030 - .0056 - .0002 - .0003 - \cdots = .3352.$$

We may, therefore, take $r_t = .335$. Then $.344 + \varepsilon = .335$ and $\varepsilon = -.009$.

In a manner similar to the above an epsilon table can be set up for values of r_ϕ, and convenient ranges of proportions in the two variables. In setting up these tables we clearly have the direction of either variable at our disposal.

We use this freedom to restrict one set of proportions to be less than .5, and to insist that r_ϕ be positive. The second variable will then have proportions ranging from 0 to 1. If in any application a negative r_ϕ occurs it is only necessary to consider the corresponding positive value in relation to the complementary proportion $1 - p$. (See Table 4).

The use of the tables is shown by the following examples. Assume that for a given 4-cell table $r_\phi = .6$, and that the corresponding pro-

*For later reference Pearson's form of this infinite series is

$$\frac{d}{N} = \tau_0(h)\tau_0(k) + \tau_1(h)\tau_1(k)r_t + \tau_2(h)\tau_2(k)r_t^2 + \tau_3(h)\tau_3(k)r_t^3 + \cdots$$

portions in the variables are .3 and .5. Entering the section of the tables corresponding to $r_\phi = .6$ with these proportions $\varepsilon = -.02$. Hence $r_t = .58$.

For a second set of data assume $r_\phi = -.8$, $p = .3$, $p' = .7$. This is equivalent to using $r_\phi = .8$, $p = .3$, and $p' = .3$. Thus $\varepsilon = -.12$, and tabular $r_t = .68$. Hence, for the data $r_t = -.68$.

For values of r_ϕ nearer 1 than .6, the iteration process described becomes impractical as the method must be repeated more and more often with twenty or more terms of the series becoming significant. For these larger values, then, we make use of Pearson's Bivariate Tables (4), employing a graphical procedure to obtain results accurate to one significant figure.

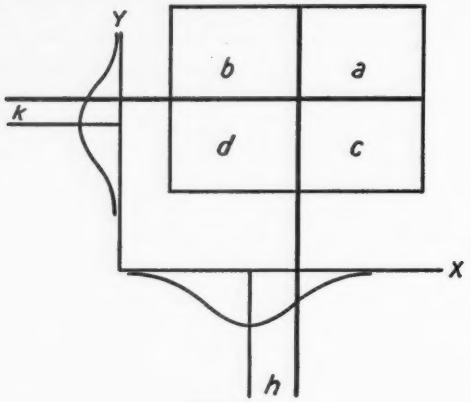


FIGURE 1

Pearson's tables supply the quantity d/N for specified h, k, and r_t as shown in Fig. 1. Specifically, h and k run by tenths from 0 to 2.6 and r_t runs by intervals of .05 from 1.00 to —1.00. Since, in Pearson's notation, $r_\phi = \dfrac{d/N - \tau_0(h)\ \tau_0(k)}{\tau_1(h)\ \tau_1(k)}$ it is possible to graph r_ϕ as a function of chosen values of r_t for fixed h and k. Then, for specified r_ϕ the graph can be read in reverse to yield r_t; a process which yields $\varepsilon = r_t - r_\phi$.

We now illustrate the graphical procedure used for the special case $h = k = 0$; that is, for proportions of .5 in both variables. In particular, then, Pearson's equation becomes

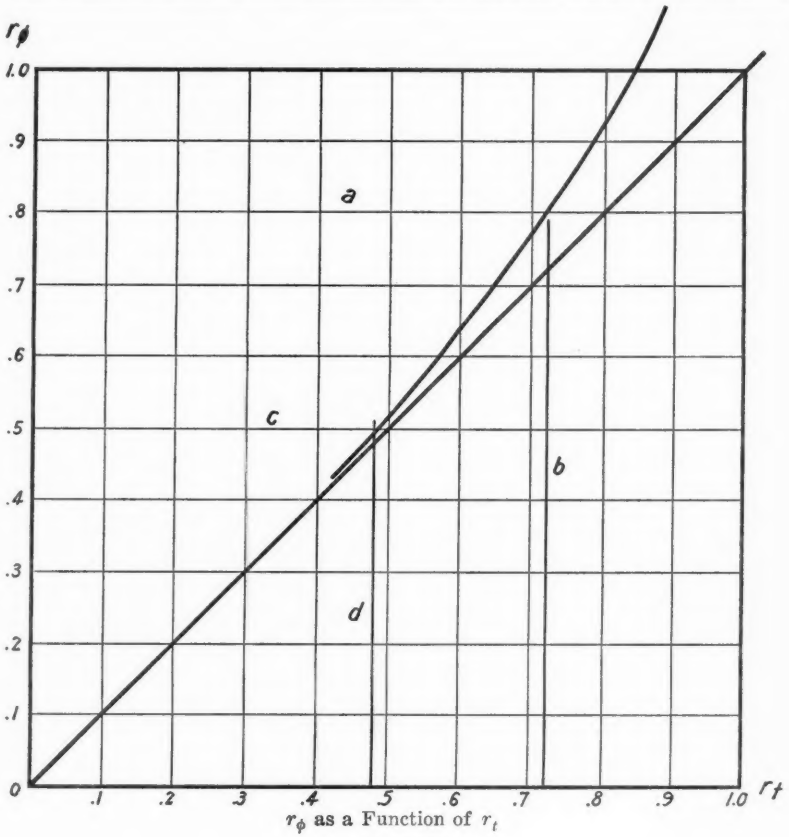
$$r_\phi = \frac{d/N - \tau_0(0)\ \tau_0(0)}{[\tau_1(0)]^2}$$

for which $[\tau_1(0)]^2 = .1592$ and $[\tau_0(0)]^2 = .25$. Hence, $r_\phi = 6.28\ d/N - 1.571$. With the aid of the latter equation and Pearson's tables we can now set up the desired correspondence between chosen r_t, and r_ϕ in Table 3.

TABLE 3

r_ϕ as a Function of r_t for $p = p' = .5$

r_t	.1	.2	.3	.4	.5	.6	.7	.8	.9
d/N	.266	.282	.298	.315	.333	.352	.373	.398	.428
$6.28d/N$	1.671	1.772	1.876	1.982	2.09	2.21	2.35	2.50	2.69
r_ϕ	.1002	.201	.305	.412	.524	.644	.775	.927	1.120



r_ϕ as a Function of r_t

FIGURE 2

The next step in the process is to graph r_ϕ as a function of r_t as shown in Fig. 2.

As previously indicated the graph can now be used in reverse to obtain for Table 4 the r_t's, for specified r_ϕ's, which will yield the desired ε's.

For clarity, let us consider two instances of this process. From lines a and b in Fig. 2 it can be seen that if $r_\phi = .8$, then $r_t = .72$. Hence $\varepsilon = -.08$. As a second illustration let $r_\phi = .5$. From lines c and d in Fig. 2, $r_t = .48$. Hence $\varepsilon = 48 - .50 = -.02$.

It is interesting to note that the graphical method becomes inaccurate for small ε, exactly the situation where iteration works best. Thus the two methods complement each other very well with excellent checks over the middle range of r_ϕ.

Examination of Table 4 will show the limitations to the estimation of tetrachoric r by means of formulas (4) and (6). Only for small and moderate values of r can reasonably accurate estimates be made, and only when p and p' are not extreme. The many vacant cells in the table indicate that estimates for higher values of r and for the more extreme values of p and p' are determinate. Some of the larger correction values that are near those vacant areas should be used with hesitation. In spite of these limitations, however, there is much utility in the complete procedure given here. Most of the correlations obtained in the social sciences are moderate or low. The wise investigator also attempts to make divisions near the medians when he resorts to artifical dichotomies. Conditions are therefore, usually favorable for utilizing these estimation procedures well within the limitations mentioned.

Illustrative Examples

In the interest of clarification we now present specific examples of the general procedures which have been developed.

The phi coefficient for the data of Table II can be obtained from the standard formula (3, p. 341):—

$$\phi = \frac{ad - bc}{[(a+b)(a+c)(b+d)(c+d)]^{\frac{1}{2}}}.$$

Substituting and extracting square roots,

$$\phi = \frac{44860}{(23.26)(23.66)(19.24)(19.72)} = .215.$$

By use of formula (4),

TABLE 4

Values of ε for Different Combinations of r_ϕ, p, and p'*

p'										
.9	.8	.7	.6	.5	.4	.3	.2	.1	p	r_ϕ
.00	.01	.00	—.01	.00	.00	.00	.00	.00	.5	
.02	.01	.00	.00	.00	.00	.00	—.01	.01	.4	
.00	.00	.01	.00	.00	.00	—.01	—.01	—.01	.3	.2
.03	.02	.00	.01	.00	—.01	—.01	—.01	—.02	.2	
.02	.02	.01	.01	.00	.01	—.01	—.02	—.02	.1	
.02	—.01	—.01	—.01	—.01	—.01	—.01	.00	.01	.5	
.03	.01	.01	.00	—.01	—.02	—.02	—.02	—.03	.4	
.06	.05	.02	.04	—.01	—.02	—.03	—.04	—.03	.3	.4
.2	.2	.04	.01	.00	—.02	—.04	—.05	—.06	.2	
.2	.2	.09	.04	.01	—.03	—.03	—.06	—.08	.1	
—.01	.00	—.01	—.02	—.02	—.02	—.02	—.01	—.02	.5	
.08	.02	.00	—.03	—.02	—.03	—.03	—.03	—.05	.4	
.1	.1	.03	.01	—.02	—.03	—.05	—.06	—.04	.3	.5
	.2	.09	.03	—.01	—.03	—.06	—.07	—.08	.2	
		.2	.08	.02	—.05	—.04	—.08	—.1	.1	
.04	.01	—.03	—.04	—.03	—.03	—.02	—.01	.04	.5	
.1	.06	.01	—.03	—.03	—.05	—.05	—.04	—.07	.4	
	.1	.06	.01	—.02	—.05	—.07	—.07	—.08	.3	.6
		.2	.05	—.01	—.04	—.07	—.09	—1.	.2	
			.1	.04	—.07	—.08	—.1	—.2	.1	
.1	.00	—.03	—.04	—.05	—.05	—.04	—.02	.07	.5	
	.1	.01	—.04	—.05	—.07	—.07	—.06	.00	.4	
		.1	.01	—.04	—.07	—.1	—.1	—.08	.3	.7
			.1	—.02	—.06	—.1	—.1	—.2	.2	
				.07	.00	—.08	—.2	—.2	.1	
	.01	—.05	—.07	—.08	—.08	—.06	.00		.5	
		.01	—.06	—.08	—.09	—.09	—.07	—.02	.4	
			.02	—.06	—.09	—.1	—.1	—.1	.3	.8
					—.07	—.1	—.2	—.2	.2	
					—.02	—.1	—.2	—.3	.1	
		—.06	—.09	—.09	—.09	—.07	.00		.5	
		.01	—.06	—.09	—.1	—.1	—.09		.4	
				—.07	—.1	—.1	—.1	—.1	.3	.85
					—.09	—.1	—.2	—.2	.2	
						—.1	—.2	—.2	.1	
		—.09	—.1	—.1	—.1	—.07	—.02		.5	
			—.08	—.1	—.1	—.1	—.1		.4	
				—.07	—.1	—.1	—.2	—.1	.3	.9
				—.02	—.1	—.2	—.2	—.2	.2	
						—.1	—.2	—.3	.1	

*$\varepsilon = r_t - r_\phi$, hence $r_t = r_\phi + \varepsilon$.

$$r_\phi = (.215)\,(1.263)\,(1.269) = .345\,.$$

This checks well with the r_ϕ of .344 obtained for the same data on page 341 through first-order use of the tetrachoric series.

By use of Table 4 and interpolation, $\varepsilon = -.01 = r_t - r_\phi$. Hence, $r_t = .344 - .01 = .334$. This checks well with the r_t of .335 obtained on page 341 through iterative use of the tetrachoric infinite series.

REFERENCES

1. Camp, B. H. The mathematical part of elementary statistics. Boston: Heath & Co., 1931.
2. Edwards, A. L. Statistical analysis. New York: Rinehart, 1946.
3. Guilford, J. P. Fundamental statistics in psychology and education, Second Ed. New York: McGraw-Hill Book Co., 1950.
4. Pearson, Karl. Tables for statisticians & biometricians. England: Cambridge University Press, 1931.
5. Peatman, J. G. Descriptive and sampling statistics. New York: Harper & Bros., 1947.
6. Peters, C. C., and Van Voorhis, W. R. Statistical procedures and their mathematical bases. New York: McGraw-Hill Book Co., 1940.

Manuscript received 10/5/50.
Revised manuscript received 3/8/51.

PSYCHOMETRIKA—VOL. 16, NO. 3
SEPTEMBER, 1951

BOOK REVIEWS

JOHN D. TRIMMER. *Response of Physical Systems.* New York: John Wiley and Sons, Inc., 1950, pp. ix + 268.

Although this book is written primarily for workers in engineering and applied physics, the author hopes that it may be of use to workers in other sciences. It is only in the first chapter that application is made to any extent to problems other than those of applied physics. But the point of view which is developed by the author, together with the mathematical results presented, would make it worthwhile for the biologist or psychologist who is interested in formulating his own problems in a quantitative, theoretical manner, to make use of this book. The text is written in such a way that only a limited knowledge of general physics and of differential equations is required in order to follow the principal ideas developed.

In the first chapter, "A Pattern for Systems," the problem to be treated is outlined. The author considers a "system" (e.g., apparatus, organism, mind, society) upon which there is a "forcing" (force, drive, motive, stimulus) which acts according to some "law" (equation, habit) to produce a "response" (action, output, mood, thought). The types of problems one may wish to solve can then be classified by the unknown. Thus, the direct problem is to find the response from the known forcing, system, and law. The converse problem is to find the forcing, the inverse problem is to find the properties of the system, and the inductive problem is to find the law. In this chapter and elsewhere the difficulties of isolating the system and its parts are brought out. In the second chapter, "Physical Systems," is given a general discussion and a classification of systems together with definitions of terms.

In chapters three, four, and six (First-Order, Second-Order, and Higher-Order Systems) a number of problems are solved. Among these are thermal, electrical, and hydraulic models, automatic speed control, production of radio-isotopes, and the nuclear chain reactor. The stability of a response and transient responses are also discussed in detail.

Chapter five, "Sinusoidal Forcing of Linear Systems," gives the solution to a second-order system with a sinusoidal input.

Chapter seven, "Measuring Instruments," treats instruments as physical systems. Such problems as range, efficiency, and accuracy are discussed. Considerable attention is given to errors—mechanism, scale, environmental, dynamic, reading, measurement, determinate, sampling, and random errors.

In chapter eight, "Feedback Systems," thermal and electronic examples are used as illustrations of direct feedback, e.g., the response itself can directly change the magnitude of the forcing. An example of negative impedance is also given. In chapter nine, "Parametric Forcing," the system is considered in which the response is altered by a change in the value of some parameter. Examples used are the condenser microphone and the nuclear chain reactor. If the response is able to alter the value of some parameter, the result is a parametric feedback system.

In chapter ten, "Distributed Systems," problems are treated in which the values of one or more of the properties of the system are distributed continuously in space in such a manner that this fact cannot be ignored. This leads to partial differential equations which are briefly discussed. Applications are made to wave motion and to the cable problem.

The last chapter, "Nonlinear Systems," treats principally the problem of a charged particle in a force field and the problem of the oscillator.

In addition to an appendix of problems, there is an appendix on the use of the Laplace transform in the study of linear systems.

The University of Chicago *H. D. Landahl*

PALMER O. JOHNSON. *Statistical Methods in Research.* New York: Prentice-Hall, Inc., 1949, pp. xvii + 377.

The reviewer wishes to congratulate Professor Johnson on the production of a statistical text that sets a new benchmark of merit in the field of statistical psychology. The author aimed to write a non-mathematical but essentially accurate presentation based on the contributions of such statisticians as R. A. Fisher, E. S. Pearson, J. Neyman, S. S. Wilks, H. Hotelling, etc. The large set of references to original articles bears witness to the diligence with which he has culled the statistical field for theories and techniques which would be of use to the research worker in psychology or education. How well he has succeeded can best be judged by examining the topics covered.

The book opens with a short discussion on the realm of statistics in which the author emphasizes that the student must be taught (1) how to choose the most effective statistical tool for the purpose in mind (2) the basic assumptions underlying the statistical tool selected (3) how to test whether these basic assumptions are fulfilled by the particular situation to which the tool is applied. The author adheres to this viewpoint rigorously in the greater part of the book; for example, all analysis of variance examples give as their first step, the testing of the assumption of homogeneity of variance.

A short discussion of 'Probability and Likelihood' which follows contains a brief account of Bayes Theorem and maximum likelihood. Maximum-likelihood estimates are used with great frequency in the latter parts of the book.

The next chapter, on 'Sampling Distributions,' gives us the sampling distributions of such frequently used statistics as the mean; the variance, Student's t; the correlation coefficient; Fisher's z ratio; the binomial distribution; etc. This chapter characterizes the approach of the author to advanced problems in mathematical statistics. Exact formulas are given, but not derived if unfamiliar mathematical concepts like Gamma functions are involved, or mathematics beyond differential calculus is necessary. All of these sampling distributions are used in a later chapter which summarizes the simple useful tests of significance. These chapters include some relatively unknown tests such as the L_1 for the homogeneity of variances and Hotelling's test for the equality of two correlation coefficients from the same sample. It is of interest to note that Hotelling's test can be generalized to test for the equality of correlation coefficients in any column of a correlation matrix.

A short explanation of the Neyman-Pearson theory of testing statistical hypotheses precedes the chapter giving examples of tests of significance. The

modern statistical theory of inference is competently summarized in terms of the concepts of the null hypothesis, Type I and Type II errors, critical regions, and the power of the likelihood ratio. In accordance with the author's principles, the presentation is on a verbal level, relying on logical formulation rather than rigorous mathematical derivations. Illustrative examples and empirical demonstrations are given at many points in the exposition. The mathematical symbols are used mainly for clear, concise, and correct formulations of these concepts.

The next chapter deals with the techniques of estimating population parameters by means of statistics. The maximum-likelihood method is used for point estimation; both Fisher's fiducial method and the Neyman-Pearson method of confidence belts are used for interval estimation. Maximum-likelihood methods are applied to such novel problems as Jackson's estimation of the reliability of a test, and Wilks' test for the equivalence of two forms of a test with respect to means, variances, and covariances. The treatment of reliability follows Jackson and Hoyt in using the analysis of variance assumptions of additivity of components and equal error variances. To the reviewer, the analysis of variance probability model seems distinctly inferior to the regression probability model for the estimation of true and error variance. Also, the assumption of homogeneity of the observed variances is unnecessarily restrictive.

The next section, on testing the hypothesis of normality, includes some useful advice on how to normalize the set of observations in those cases where the distributions can be shown to be certain non-normal types. Then there is a unique chapter on those tests of significance which are 'distribution-free' or 'nonparametric,' i.e., which make no assumptions about the distribution of the observations in the parent population. For most psychologists, the analysis of variance using ranked data, Kendall's W coefficient of concordance, and the use of paired comparisons for calculating the agreement and consistency of subjects, will be new techniques. This chapter is a good idea and the reviewer hopes it will be expanded in the next edition to take in Kendall's rank-order coefficient *tau*, simple devices like Tchebycheff's inequality, and G. W. Brown's analysis of variance using medians instead of means.

A discussion of sampling is followed by a detailed mathematical presentation of the analysis of variance. The probability models for single classification and double classification (rows by columns) are set up, and the maximum-likelihood derivation is given. (Incidentally, neither here nor elsewhere is it explicitly pointed out that the maximum-likelihood method is equivalent to the least-squares technique only because of the assumption of normality.) A number of examples of the analysis of variance and covariance are presented.

Two chapters are devoted to the principles and applications of experimental design with examples of experiments in psychology using randomized blocks, Latin squares, and factorial designs. It is gratifying to note that the probability model for each example is given in full for all applications in analysis of variance and covariance. This method of teaching Fisher's techniques far surpasses Fisher's own presentation of his methods. By explicitly designating which parameters are being tested for their deviation from zero, it allows the student to do his own reasoning instead of forcing him to follow some blind, rule-of-thumb handbook. To leave out the probability models, as most textbooks do, is to ensure that a certain proportion of students will go thru the ritual of analysis of vari-

ance with the religious conscientiousness of a devotee of number magic, and for the same mystical reasons.

The book concludes with a succinct summary of multiple regression problems. Unfortunately Fisher's computational technique is used, which means that the complete inverse of the correlation matrix is computed each time. Even worse, it is not stated explicitly anywhere that it is the inverse of a matrix which is being computed. Tests of significance for the multiple correlation and the multiple regression weights are given and illustrated. The two-group linear discriminant function is derived using Mahalanobis's D^2. The relation of D^2 to Hotelling's T^2 is mentioned but not explained in detail. Wherry's proof that Fisher's linear discriminant function is exactly proportional to the regression of the dichotomous criterion of group membership on the independent variables is not mentioned. Consequently, it is not pointed out that testing D^2 or T^2 for significance is algebraically identical to testing a multiple correlation for significance.

The book is designed for a year's course of advanced statistics, and assumes that the student will have a knowledge of descriptive statistics and elementary calculus. The reviewer has been able to use the chapters on the testing of statistical hypotheses as the basis of the third term in a year's course in statistics. The students were clinical psychologists and for the most part knew no calculus. Yet, with Johnson's treatment, it was possible to give a firm rational foundation to all tests of significance by using such concepts as maximum likelihood, null hypothesis, Type I and Type II errors, critical region, etc.

However, the instructor will probably find the book of more use than will the student. Johnson has undertaken the arduous task of listing every reference that he has used. We thus have a compilation of almost all those articles which would be of interest to a statistical psychologist both from a theoretical and applied point of view.

There are several points on which the reviewer finds himself in disagreement with the author. The difference between a one-tail and a two-tail test is never made explicit and apparent errors occur in the applications. For example in Problem V.1 (pp. 69-70) the question is asked 'if the mean ability of the class is the same as that of the population.' This question seems to call for a two-tail test since the direction of the difference is not specified. But a one-tail test is used. Johnson has stated (in a personal communication to the reviewer) that the one-tail test must be used because the population mean is known in this problem. This does not seem to be adequate. If the one-tail test were used on each such case at, let us say, the 5% level, then, when the null hypothesis is true, 10% of the differences would be rejected as non-null differences. It seems to the reviewer that the Neyman-Pearson theory demands that the level of confidence correspond exactly to the per cent of Type I errors.

A similar case arises when the F ratio is used to test the homogeneity of two variances (p. 82). The tabled value of the F ratio gives a one-tail test. But since the direction of the difference is not specified, a two-tail test is necessary. We must find the probability that an F ratio would exceed s^2_1/s^2_2 or be less than s^2_2/s^2_1 (where $s^2_1 > s^2_2$). The L_1 test for homogeneity of the two variances gives this same two-tail F test. A. M. Mood demonstrates the point very clearly in his recent text (*Introduction to the Theory of Statistics.* New York: McGraw-Hill Book Co., 1950, p. 268).

The treatment of chi-square is utterly inadequate. This is all the more sur-

prising in view of the care taken to define other basic sampling distributions, such as t and Fisher's z. The first mention of chi-square comes on page 37 in the middle of a discussion on the goodness of fit of sampled means to a known normal distribution. The inquiring student is referred to page 96 where he will find an example of the use of chi-square, but not a word of its meaning, or the basic sampling distribution it illustrates. But as Lewis and Burke have pointed out, statistical textbooks for psychologists seem to universally fail when the chi-square test is discussed (D. Lewis and C. V. Burke, *Psychol. Bull.*, 1949, **46**, 433-489). Their article, however, should go a long way towards rectifying this deficiency.

In spite of these defects, the book remains unexcelled in its field. Author Johnson has managed to pack more good statistics into his book than appears in any other comparable text. *Statistical Methods in Research* is definitely required reading for all teachers of statistics in the fields of psychology and education.

University College, London *Ardie Lubin*

HOWARD W. GOHEEN and SAMUEL KAVRUCK, *Selected References on Test Construction, Mental Test Theory, and Statistics, 1929-1949*, Washington, D. C.: United States Civil Service Commission, 1950, pp. 209.

This is an exceedingly valuable index of much of the literature in the fields of test theory, test construction, and statistics. Recommended for all members of the Society.

University of Michigan *Clyde H. Coombs*